

# Optimizing the placement of tap positions and guess and determine cryptanalysis with variable sampling

S. Hodžić, E. Pasalic, and Y. Wei<sup>\*†</sup>

## Abstract

<sup>1</sup> In this article an optimal selection of tap positions for certain LFSR-based encryption schemes is investigated from both design and cryptanalytic perspective. Two novel algorithms towards an optimal selection of tap positions are given which can be satisfactorily used to provide (sub)optimal resistance to some generic cryptanalytic techniques applicable to these schemes. It is demonstrated that certain real-life ciphers (e.g. SOBER-t32, SFINKS and Grain-128), employing some standard criteria for tap selection such as the concept of full difference set, are not fully optimized with respect to these attacks. These standard design criteria are quite insufficient and the proposed algorithms appear to be the only generic method for the purpose of (sub)optimal selection of tap positions. We also extend the framework of a generic cryptanalytic method called Generalized Filter State Guessing Attacks (GFSGA), introduced in [26] as a generalization of the FSGA method, by applying a variable sampling of the keystream bits in order to retrieve as much information about the secret state bits as possible. Two different modes that use a variable sampling of keystream blocks are presented and it is shown that in many cases these modes may outperform the standard GFSGA mode. We also demonstrate the possibility of employing GFSGA-like attacks to other design strategies such as NFSR-based ciphers (Grain family for instance) and filter generators outputting a single bit each time the cipher is clocked. In particular, when the latter scenario is considered, the idea of combining GFSGA technique and algebraic attacks appears to be a promising unified cryptanalytic method against NFSR-based stream ciphers.

**Index terms**— Stream ciphers, Filter generators, Generalized filter state guessing attacks, Tap positions, Algebraic attacks.

## 1 Introduction

Certain hardware oriented symmetric-key encryption schemes employ two basic primitives for generating keystream sequences, namely a linear feedback shift register (LFSR) and a nonlinear Boolean function. The LFSR is mainly used for storing the state bits and for providing the

---

<sup>\*</sup>S. Hodžić and E. Pasalic are with the University of Primorska, FAMNIT, Koper, Slovenia (e-mails: samir.hodzic@famnit.upr.si, enes.pasalic6@gmail.com).

<sup>†</sup>Y. Wei is with the Guilin University of Electronic Technology, Guilin, P.R. China (e-mail: walker\_wei@msn.com).

<sup>1</sup>A part of this paper, related to algorithms for finding an optimal placement of tap positions, was presented at the First BalkanCryptSec conference, Istanbul, Turkey, October 2014.

inputs to the Boolean function, which in turn processes these inputs to output a single bit as a part of the keystream sequence. For a greater throughput, a vectorial Boolean function, say  $F$ , may be used instead to provide several output bits at the time, thus  $F : GF(2)^n \rightarrow GF(2)^m$ . Nonlinear filter generator is a typical representative of a hardware oriented design in stream ciphers. It consists of a single linear feedback shift register (LFSR) and a nonlinear function  $F : GF(2)^n \rightarrow GF(2)^m$  that processes a fixed subset of  $n$  stages of the LFSR. This fixed subset of the LFSR's cells is usually called the *taps*.

The resistance of nonlinear filter generators against various cryptanalytic attacks, such as (fast) correlation attacks [16, 25, 19], algebraic attacks [5, 6, 17], probabilistic algebraic attacks [4, 23], attacks that take the advantage of the normality of Boolean functions [20] etc., mainly depends on the choice of the filtering function  $F$ . The design rules for ensuring good security margins against these attacks are more or less known today. Nevertheless, guess and determine cryptanalysis is a powerful cryptanalytic tool against these schemes whose efficiency is irrelevant of the cryptographic properties of the filtering function (the same applies to time-memory-data trade-off attacks [3], [12], [13]) but rather to the selection of LFSR: its size, primitive polynomial used and tapping sequence (tap positions used to supply  $F$  with the inputs). The main goal of the guess and determine cryptanalysis, when applied to these schemes, is to recover the secret state bits contained in the LFSR by guessing a certain portion of these bits and exploiting the structure of the cipher. The term structure here mainly refers to the tap positions of LFSR used to provide the inputs to  $F$  and to some fixed positions of LFSR for implementing a linear recursion through the primitive connection polynomial. For the first time, it was explicitly stated in [9] that the choice of tap sequence may play more significant role than the optimization of  $F$  in the context of inversion attacks introduced in [9], see also [8, 10]. To protect the cipher from inversion attacks a full positive difference set was employed in [9], where a set of positive integers  $\Gamma = \{i_1, \dots, i_n\}$  is called a full positive difference set if all the positive pairwise differences between its elements are distinct. These sets are, for instance, used in the design of self-orthogonal convolutional codes.

The basic idea behind the attacks similar to inversion attacks is to exploit the shift of the secret state bits that are used as the input to the filtering function. To understand this assume that we sample the keystream bits at suitable time instances so that a portion of the secret bits that are used in the previous sampling instance appear again as (part of) the input (at different positions) to the filtering function. Provided the knowledge of the output, this information then significantly reduces the uncertainty about the remaining unknown inputs. The designers, well aware of the fact that a proper tap selection plays an important role in the design, mainly use some standard (heuristic) design rationales such as taking the differences between the positions to be prime numbers (if possible), the taps are distributed over the whole LFSR etc.. Intuitively, selecting the taps at some consecutive positions of the LFSR should be avoided (see also [1]), and similarly placing these taps at the positions used for the realization of the feedback connection polynomial is not a good idea either.

Even though a full positive difference set is a useful design criterion which ensures that there are no repetitions of several input bits, it is quite insufficient criterion which does not prevent from the attacks such as GFSGA (Generalized Filter State Guessing Attack) introduced in [26]. For instance, assume for simplicity that the inputs to the filtering function are taken at tap positions  $\mathcal{I} = \{3, 6, 12, 24\}$  of the employed LFSR, thus our filtering function takes four inputs,

i.e.,  $n = 4$ . It is easily verified that all the differences are distinct and the set of (all possible) differences is  $D^{\mathcal{I}} = \{i_j - i_k : i_j, i_k \in \mathcal{I}, i_j > i_k\} = \{3, 6, 9, 12, 18, 21\}$ . Nevertheless, all these numbers being multiple of 3 would enable an efficient application of GFSGA-like cryptanalysis since the information about the previous states would be maximized.

Another criterion considered in the literature, aims at ensuring that a multiset of differences of the tap positions is mutually coprime. This means, that for a given set of tap positions  $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$  of an LFSR of length  $L$  (thus  $1 \leq i_1 < i_2 < \dots < i_n \leq L$ ) all the elements in the difference set  $\mathcal{D}^{\mathcal{I}} = \{i_j - i_k : i_j, i_l \in \mathcal{I}, i_j > i_k\}$  are mutually coprime. This condition, which would imply an optimal resistance to GFSGA-like methods, is easily verified to be impossible to satisfy (taking any two odd numbers their difference being even would prevent from taking even numbers etc.). Therefore, only the condition that the consecutive distances are coprime appears to be reasonable, that is, the elements of  $D = \{i_{j+1} - i_j : i_j \in \mathcal{I}\}$  are mutually coprime. An exhaustive search is clearly infeasible, since in real-life applications to select (say)  $n = 20$  tap positions for a driving LFSR of length 256 would give  $\binom{256}{20} = 2^{98}$  possibilities to test for optimality.

This important issue of finding (sub)optimal solutions for selecting tap positions, given the input size  $n$  and the length  $L$  of the driving LFSR, appears to be highly neglected in the literature. Certain criteria for the choice of tap positions was firstly mentioned in [9] but a more comprehensive treatment of this issue was firstly addressed recently in [22], where two algorithms for the purpose of selecting taps (sub)optimally were presented. These algorithms were used in [22] to show that the selection of tap positions in real-life stream ciphers such as SOBER-t32 [11] and SFINKS [27] could have been (slightly) improved to ensure a better resistance of these ciphers to GFSGA-like cryptanalysis. For self-completeness and since this manuscript extends the work in [22], the above mentioned algorithms and the theoretical discussion for their derivation is also given, though in a slightly suppressed form. The emphasis is given to the construction of algorithms and their relation to the criteria for tap selection proposed in [9]. It is shown that these criteria are embedded in our algorithms but they are not sufficient for protecting the considered encryption schemes against GFSGA-like methods adequately. In particular, the selection of tap positions for Grain-128 cipher is far from being optimized allowing for a significant improvement of its resistance to GFSGA-like attacks as shown in Section 5. Thus, these algorithms appear to be the only known efficient and generic method for the purpose of selecting tap positions (sub)optimally.

Another goal of this manuscript is to further extend the GFSGA framework by considering a variable mode of sampling which was not addressed in FSGA [24] or GFSGA [26]. For a better understanding of the differences between various modes we give a brief description of the known modes, thus FSGA and GFSGA, and discuss our extended mode of GFSGA which we describe later in more detail. The basic idea behind FSGA is to recover secret state bits by reducing the preimage space of the filtering function  $F : GF(2)^n \rightarrow GF(2)^m$  using the knowledge of previously guessed bits. Since for uniformly distributed  $F$  there are  $2^{n-m}$  preimages for any observed  $m$ -bit output block, the attacker may for each choice of  $2^{n-m}$  many possible inputs (over the whole set of sampling instances) set up an overdefined system of linear equations in secret state bits. This attack turns out to be successful only for relatively large  $m$ , more precisely for approximately  $m > n/2$ . In certain cases, the running time of FSGA may be lower than the running time of a classical algebraic attack (cf. [24]). Nevertheless, the placement of tap

positions of a nonlinear filter generator were of no importance for this attack. More precisely, only one bit of the information was considered to be known from the previous states in the case of FSGA. The complexity of the attack was significantly improved in [26], where the information from the neighbouring taps, in the attack named GFSGA (Generalized FSGA), was used for a further reduction of the preimage space. In particular, the attack complexity of GFSGA is very sensitive to the placement of taps, which essentially motivated the initiative taken in [22] for devising the algorithms for computing (sub)optimal choices of tap positions that give the maximum resistance against GFSGA.

However, even this generalized approach, which takes into account the tap positions of the driving LFSR, turns out not to be fully optimal. The main reason is that GFSGA works with a constant sampling rate and its optimal sampling rate can be easily computed given the set of tap positions. The algorithms themselves, as presented in [22], are designed to select tap positions that gives the largest resistance to GFSGA with a constant sampling rate. Therefore, a natural question that arises here regards the impact on the robustness of the cipher if a variable sampling rate is used instead. This issue is elaborated here by introducing two different modes of GFSGA with non-constant (variable) sampling rate. These modes are much less dependent on the choice of tap positions and in many cases their performance is demonstrated through examples to be better than that of the standard GFSGA.

We notice that the main difficulty, when comparing the performance of these modes theoretically, lies in the fact that there are intrinsic trade-offs between the main parameters involved in the complexity computation, cf. Remark 4. The main reason is that each of these modes attempt to reduce the preimage space based on the knowledge of some secret state bits that reappear as the inputs, but at the same time these linear equations (describing the known/guessed secret state bits) have already been used for setting up a system of linear equations to be solved once the system becomes overdefined. Thus, increasing the number of repeated bits makes a reduction of the preimage space more significant (less bits needs to be guessed) but at the same time more sampling is required since the repeated bits do not increase the rank of the system of linear equations. This is the trade-off that makes the complexity analysis hard and consequently no theoretical results regarding the performance of the attack modes can be given.

Finally, well aware of the main limitation of GFSGA-like attacks, which are efficiently applicable to LFSR-based ciphers with filtering function  $F : GF(2)^n \rightarrow GF(2)^m$  where  $m > 1$ , we briefly discuss their application to single output filtering functions (thus  $m = 1$ ) and to ciphers employing nonlinear feedback shift registers (NFSRs). In both cases we indicate that GFSGA attacks may be adjusted to work satisfactory in these scenarios as well. Most notably, there might be a great potential in applying GFSGA attacks in combination with other cryptanalytic techniques such as algebraic attacks. This possibility arises naturally due to the fact that GFSGA-like attacks reduce the preimage space of possible inputs to a filtering function using the knowledge of previous inputs, thus giving rise to the existence of low degree annihilators defined on a restriction of the filtering function (obtained by keeping fixed a subset of known input variables). In another direction, when considering NFSR-based ciphers we propose a novel approach of mounting internal state recovery attacks on these schemes which employs the GFSGA sampling procedure but without solving the deduced systems of equations at all. More precisely, this new type of internal state recovery attack collects the outputs within a certain sampling window which then enables an efficient recovery of a certain portion of internal state

bits. This is done by filtering out the wrong candidates based on the knowledge of reduced preimage spaces that correspond to the observed outputs.

The rest of the article is organized as follows. In Section 2, some basic definitions regarding Boolean functions and the mathematical formalism behind the structure of nonlinear filter generators is given. For completeness, a brief overview of FSGA and GFSGA is also given in this section. Two different modes of GFSGA with variable sampling distance are introduced in Section 3. The complexity analysis of these modes in terms of the number of repeated state bit equations is addressed here as well. In Section 4, the performance of different attack modes and the algorithms for determining a (sub)optimal selection of tap positions are presented. The possibility of applying GFSGA to single output filtering functions and to NFSR-based ciphers is discussed in Section 5. Some concluding remarks are given in Section 6.

## 2 Preliminaries

A Boolean function is a mapping from  $GF(2)^n$  to  $GF(2)$ , where  $GF(2)$  denotes the binary Galois field and  $GF(2)^n$  is an  $n$ -dimensional vector space spanned over  $GF(2)$ . A function  $f : GF(2)^n \rightarrow GF(2)$  is commonly represented using its associated algebraic normal form (ANF) as follows:

$$f(x_1, \dots, x_n) = \sum_{u \in GF(2)^n} \lambda_u \left( \prod_{i=1}^n x_i^{u_i} \right),$$

where  $x_i \in GF(2)$ ,  $(i = 1, \dots, n)$ ,  $\lambda_u \in GF(2)$ ,  $u = (u_1, \dots, u_n) \in GF(2)^n$ . A vectorial (multiple output) Boolean function  $F(x)$  is a mapping from  $GF(2)^n$  to  $GF(2)^m$ , with  $m \geq 1$ , which can also be regarded as a collection of  $m$  Boolean functions, i.e.,  $F(x) = (f_1(x), \dots, f_m(x))$ . Commonly,  $F(x)$  is chosen to be uniformly distributed, that is,  $\#\{x \in GF(2)^n \mid F(x) = z\} = 2^{n-m}$ , for all  $z \in GF(2)^m$ . In some cases, we also use the notation  $|A|$  to denote the cardinality of the set  $A$ . Moreover, for any  $z = (z_1, \dots, z_m) \in GF(2)^m$ , we denote the set of preimage values by  $S_z = \{x \in GF(2)^n \mid F(x) = z\}$ .

### 2.1 Nonlinear filter generator

A nonlinear filter generator [18] consists of a single LFSR of length  $L$  (thus comprising  $L$  memory cells) whose  $n$  fixed positions (taps) are used as the inputs to a filtering function  $F : GF(2)^n \rightarrow GF(2)^m$ . These  $m$  outputs are used as a part of the keystream bits each time the LFSR is clocked and the internal state is then updated by a transition function. More formally,

$$(z_1^t, \dots, z_m^t) = (f_1(\ell_n(\mathbf{s}^t)), \dots, f_m(\ell_n(\mathbf{s}^t))),$$

where  $\mathbf{s}^t = (s_0^t, \dots, s_{L-1}^t)$  is the secret state of the LFSR at time  $t$ . The notation  $\ell_n(\mathbf{s}^t)$  means that only a fixed subset of  $n$  bits of  $\mathbf{s}^t = (s_0^t, \dots, s_{L-1}^t)$  is used as the input to Boolean functions  $f_1, \dots, f_m$ , and  $z_1^t, \dots, z_m^t$  are the corresponding output keystream bits. These fixed  $n$  positions of LFSR, used as the input to  $F$ , are called *tap positions* in the sequel and will be denoted by  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\}$ .

The LFSR is updated by computing a newly generated (rightmost) bit  $s_L^{t+1}$  as a linear combination of  $s_0^t, \dots, s_{L-1}^t$  determined by the connection polynomial and then shifting its content

to the left. That is, its state is updated as  $(s_1^{t+1}, \dots, s_{L-1}^{t+1}, s_L^{t+1}) \leftarrow (s_0^t, s_1^t, \dots, s_{L-1}^t)$ . Due to linearity of its feedback connection polynomial, at any  $t \geq 0$  we have  $\ell_n(s_0^t, \dots, s_{L-1}^t) = (\psi_1^t(\mathbf{s}^0), \dots, \psi_n^t(\mathbf{s}^0))$ , where the linear functions  $\psi_i^t(\mathbf{s}^0) = \sum_{j=0}^{L-1} a_{i,j}^t s_j^0$ , ( $i = 1, \dots, n$ ), are unique linear combinations of the initial secret state bits  $\mathbf{s}^0 = (s_0^0, \dots, s_{L-1}^0)$ , at time  $t = 0$ . The binary coefficients  $a_{i,j}^t$  above can therefore be efficiently computed from the connection polynomial of LFSR for all  $t \geq 0$ .

## 2.2 An overview of FSGA and GFSGA

For self-completeness and due to the close relation with subsequent sections, we briefly describe the main ideas behind FSGA and its extension GFSGA. For both attacks there is no restriction on  $F : GF(2)^n \rightarrow GF(2)^m$ , thus  $F$  satisfies all the relevant cryptographic criteria including a uniform distribution of its preimages. This also indicates a generic nature of GFSGA and the possibility of improving its performance in case the filtering function is not optimally chosen.

## 2.3 FSGA description

For any observed keystream block  $z^t = (z_1^t, \dots, z_m^t)$  at time  $t$ , there are  $2^{n-m}$  possible inputs  $x^t \in S_{z^t}$ . Moreover, for every guessed preimage  $x^t = (x_1^t, \dots, x_n^t) \in S_{z^t}$ , one obtains  $n$  linear equations in the initial secret state bits  $s_0^0, \dots, s_{L-1}^0$  through  $x_i^t = \sum_{j=0}^{L-1} a_{i,j}^t s_j^0$ , for  $1 \leq i \leq n$ . The goal of the attacker is to recover the initial state bits  $(s_0^0, \dots, s_{L-1}^0)$  after obtaining sufficiently many keystream blocks  $z^t = (z_1^t, \dots, z_m^t)$ . If the attacker observes the outputs at some time instances  $t_1, \dots, t_c$ , so that  $nc > L$ , then with high probability each system of  $nc$  linear equations will be solvable but only one system will provide a unique and consistent (correct) solution.

There are  $2^{(n-m)c}$  possibilities of choosing  $c$  input tuples  $(x_1^{t_1}, \dots, x_n^{t_1}), \dots, (x_1^{t_c}, \dots, x_n^{t_c})$  from  $S_{z^{t_i}}$ , and for each such  $c$ -tuple a system of  $nc$  linear equations in  $L$  variables (secret state bits) is obtained. The complexity of solving a single overdefined system of linear equations with  $L$  variables is about  $L^3$  operations. Thus, the complexity of the FSGA is about  $2^{(n-m)c} L^3$  operations, where  $c \approx \lceil \frac{L}{n} \rceil$ .

## 2.4 GFSGA description

The major difference to FSGA is that the GFSGA method efficiently utilizes the tap positions of the underlying LFSR. Let the tap positions of the LFSR be specified by the set  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\}$ ,  $1 \leq l_1 < l_2 < \dots < l_n \leq L$ . If at any time instance  $t_1$ , we assume that the content of the LFSR at these tap positions is given by  $\mathbf{s}_{\mathcal{I}_0}^{t_1} = (s_{l_1}^{t_1}, \dots, s_{l_n}^{t_1})$ , then at  $t = t_1 + \sigma$  we have  $\mathbf{s}_{\mathcal{I}_0+\sigma}^{t_1+\sigma} = (s_{l_1+\sigma}^{t_1+\sigma}, \dots, s_{l_n+\sigma}^{t_1+\sigma})$ , where the notation  $\mathbf{s}_{\mathcal{I}_0}^{t_1}$  means that we only consider the state bits at tap positions. Notice that the state bits at tap positions at time instance  $t_1 + \sigma$ , denoted as  $\mathbf{s}_{\mathcal{I}_0+\sigma}^{t_1+\sigma}$ , does not necessarily intersect with  $\mathbf{s}_{\mathcal{I}_0}^{t_1} = \mathbf{s}^{t_1}$ , thus if the intersection is an empty set no information from the previous sampling can be used at  $t_1 + \sigma$ . The extreme case of a poor cryptographic design corresponds to the selection  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\} = \{l_1, l_1 + \sigma, \dots, l_{n-1} + \sigma\}$ , thus having  $l_{i+1} = l_i + \sigma$  for any  $i = 1, \dots, n-1$ . This would imply that at  $t = t_1 + \sigma$ , based on the knowledge of the state bits  $\mathbf{s}_{\mathcal{I}_0}^{t_1}$ , we would dispose with  $n-1$  input bits to  $F$  and the unknown  $n$ -th input is easily determined upon the observed output.



Nevertheless, we can always select  $\sigma$  so that at least one bit of information is conveyed. More formally, if we denote the outputs by  $z^{t_1}, \dots, z^{t_c}$  at  $t_1, \dots, t_c$ , where  $t_i = t_1 + (i-1)\sigma$  and  $1 \leq \sigma \leq (l_n - l_1)$ , the sampling process at these time instances may give rise to identical linear equations since the equations  $x_i^{t_u} = \sum_{j=0}^{L-1} a_{i,j}^{t_u} s_j$  (where  $1 \leq i \leq n$ ) may be shifted to  $x_l^{t_v} = \sum_{j=0}^{L-1} a_{i,j}^{t_v} s_j$ , for some  $1 \leq i < l \leq n, 1 \leq u < v \leq c$ . For simplicity, throughout the article the LFSR state bits at time instance  $t_i$  are denoted by  $s^i = (s_{0+i}, \dots, s_{L-1+i})$ , whereas for the state bits at tap positions we use the notation  $s^{t_i} = (s_{l_1}^{t_i}, \dots, s_{l_n}^{t_i})$ .

This mode of the GFSGA attack will be called the GFSGA with a constant sampling step, or shortly *GFSGA*. It is of importance to determine how many identical linear equations will be obtained for all the sampling instances  $t_1, \dots, t_c$ . By introducing  $k = \lfloor \frac{l_n - l_1}{\sigma} \rfloor$ , and for  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\}$  defined recursively:

$$\begin{aligned} \mathcal{I}_1 &= \mathcal{I}_0 \cap \{l_1 + \sigma, l_2 + \sigma, \dots, l_n + \sigma\}, \\ \mathcal{I}_2 &= \mathcal{I}_1 \cup \{\mathcal{I}_0 \cap \{l_1 + 2\sigma, l_2 + 2\sigma, \dots, l_n + 2\sigma\}\}, \\ &\vdots \\ \mathcal{I}_k &= \mathcal{I}_{k-1} \cup \{\mathcal{I}_0 \cap \{l_1 + k\sigma, l_2 + k\sigma, \dots, l_n + k\sigma\}\}, \end{aligned} \tag{1}$$

the analysis in [26] showed that the complexity of the *GFSGA* is closely related to the parameter  $r_i = \#\mathcal{I}_i$ , where  $i = 1, \dots, k$ .

**Remark 1** *The above notation means that if for instance some  $i \in \mathcal{I}_1$  and therefore  $i \in \mathcal{I}_0$ , then the state bit  $s_i^{t_2}$  was used in the previous sampling since it was at position  $i - \sigma \in \mathcal{I}_0$  at time  $t_1$ , where  $t_2 = t_1 + \sigma$ . The idea is easily generalized for  $\#\mathcal{I}_i = r_i$ , where  $i = 2, \dots, k$ .*

The number of identical equations obtained in [26] is given as follows. If  $c \leq k$ , then in total  $\sum_{i=1}^{c-1} r_i$  identical linear equations are obtained, whereas for  $c > k$  this number is  $\sum_{i=1}^k r_i + (c - k - 1)r_k$ . Note that in this case  $r_k = r_{k+1} = \dots = r_{c-1}$  due to the definition of  $k$ , which simply guarantees that after  $k$  sampling instances the maximum (and constant) number of repeated equations is attained. Notice that if  $c \leq k$ , then there are

$$2^{n-m} \times 2^{n-m-r_1} \times \dots \times 2^{n-m-r_{(c-1)}}$$

possibilities of choosing  $c$  input tuples  $(x_1^{t_1}, \dots, x_n^{t_1}), \dots, (x_1^{t_c}, \dots, x_n^{t_c})$ . For each such choice, a system of  $nc - \sum_{i=1}^{c-1} r_i > L$  linear equations in  $L$  state variables can be obtained, where the number of samples  $c$  ensures that the systems of equations are overdefined and there will be a unique consistent solution to these systems. Consequently, the time complexity of the attack for  $c \leq k$ , corresponding to solving  $2^{n-m} \times 2^{n-m-r_1} \times \dots \times 2^{n-m-r_{(c-1)}}$  many linear systems, was estimated as,

$$T_{Comp}^{c \leq k} = 2^{n-m} \times 2^{n-m-r_1} \times \dots \times 2^{n-m-r_{(c-1)}} \times L^3, \tag{2}$$

and similarly, if  $c > k$ , the time complexity for  $c > k$  was given by

$$T_{Comp}^{c > k} = 2^{n-m} \times 2^{n-m-r_1} \times \dots \times 2^{n-m-r_k} \times 2^{(n-m-r_k) \times (c-k-1)} \times L^3. \tag{3}$$

Notice that there is an intrinsic trade-off between the time complexity and the condition that  $nc - \sum_{i=1}^{c-1} r_i > L$  (regarding the uniqueness of solution) through the parameters  $c$  and  $r_i$ .

Indeed, the time complexity is minimized if  $c$  is minimized and  $r_i$  are maximized, but the condition  $nc - \sum_{i=1}^{c-1} r_i > L$  requires on contrary that  $c$  is maximized and  $r_i$  are minimized, see also Remark 4.

**Remark 2** *If  $n - m - r_i \leq 0$ , for some  $i \in \{1, \dots, k\}$ , then the knowledge of these  $r_i$  bits allows the attacker to uniquely identify the exact preimage value from the set of  $2^{n-m}$  possible preimages, i.e., we assume  $2^{(n-m-r_i)} = 1$  when  $n - m - r_i \leq 0$ .*

A complexity comparison of FSGA, GFSGA and CAA (Classical algebraic Attack) for certain choices of tap position was given in [26] and in certain cases the GFSGA mode of attack outperformed both other methods.

**Remark 3** *Note that when the GFSGA method is applied, the attacker chooses the sampling distance  $\sigma$  for which the minimal complexity  $T_{Comp.}$  of the attack is achieved. This distance is called an optimal sampling distance, and its calculation is done by checking the complexities of the attack for all  $\sigma \in \{1, \dots, L\}$ .*

### 3 GFSGA with a variable sampling step

In this section, we describe the GFSGA method with a variable sampling step  $\sigma$ , which we denote as  $GFSGA^*$ . The whole approach is quite similar to  $GFSGA$ , the main difference is that we consider outputs at any sampling distances, i.e., the observed outputs  $z^{t_{i-1}}$  and  $z^{t_i}$  ( $i = 1, \dots, c$ ) do not necessarily differ by a fixed constant value and consequently the sampling distances  $\sigma_1, \dots, \sigma_c$  are not necessarily the same. It turns out that this approach may give a significant reduction in complexity compared to the standard version of the attack, see Section 4.

We first adopt some notation to distinguish between the two modes. The number of observed outputs for which an overdefined system is obtained is denoted by  $c^*$ , the corresponding number of repeated bits by  $R^*$  and the attack complexity by  $T_{Comp.}^*$ . The outputs taken at time instances  $t_i$  are denoted by  $w^{t_i}$ , where we use the variable sampling steps  $\sigma_i$  so that  $t_{i+1} = t_i + \sigma_i$ , for  $i = 1, 2, \dots, c^* - 1$ . These values  $\sigma_i$  (distances between the sampled outputs), are referred to as the variable steps (distances). Throughout this article, for easier identification of repeated bits over observed LFSR states, the state bits  $s_0, s_1, \dots$  are represented via their indices in  $\mathbb{N}$ , i.e.,  $s_i \rightarrow (i + 1) \in \mathbb{N}$  ( $i \geq 0$ ). In other words, the LFSR state bits  $s^i = (s_{0+i}, s_{1+i}, \dots, s_{L-1+i})$  are treated as a set of integers given by

$$(s_{0+i}, s_{1+i}, \dots, s_{L-1+i}) \leftrightarrow \{1 + i, 2 + i, \dots, L + i\}. \quad (4)$$

The purpose of this notation is to simplify the formal definition of LFSR state bits at tap positions introduced in the previous section. In addition, it allows us to easier track these bits and to count the number of repeated bits using the standard concepts of union or intersection between the sets. Henceforth, the LFSR states  $s^i$  we symbolically write as  $s^i = \{1 + i, 2 + i, \dots, L + i\}$  ( $i \geq 0$ ), and this notation applies to state bits at tap positions  $s^{t_i}$ . For clarity, a few initial steps of the process of determining the preimage spaces is given in Appendix.



It is not difficult to see that the equalities (1), used in *GFSGA* to determine the parameters  $r_i = \#\mathcal{I}_i$ , are special case of the equalities given as:

$$\begin{aligned}
\mathcal{I}_1^* &= \mathcal{I}_0 \cap \{l_1 + \sigma_1, l_2 + \sigma_1, \dots, l_n + \sigma_1\} = s^{t_1} \cap s^{t_2}, \\
\mathcal{I}_2^* &= \{s^{t_1} \cup s^{t_2}\} \cap \{l_1 + (\sigma_1 + \sigma_2), \dots, l_n + (\sigma_1 + \sigma_2)\} = \{s^{t_1} \cup s^{t_2}\} \cap s^{t_3}, \\
&\vdots \\
\mathcal{I}_j^* &= \{s^{t_1} \cup \dots \cup s^{t_j}\} \cap \{l_1 + \sum_{i=1}^j \sigma_i, \dots, l_n + \sum_{i=1}^j \sigma_i\} = \bigcup_{i=1}^j s^{t_i} \cap s^{t_{j+1}}, \\
&\vdots \\
\mathcal{I}_{c^*-1}^* &= \bigcup_{i=1}^{c^*-1} s^{t_i} \cap s^{t_{c^*}},
\end{aligned} \tag{5}$$

where  $s^{t_1}, \dots, s^{t_{c^*}}$  represents the LFSR state bits at tap positions at time instances  $t_1, \dots, t_{c^*}$ .

In general, the number of repeated equations which corresponds to the outputs  $w^{t_1}, \dots, w^{t_{c^*}}$  at variable distances  $\sigma_i$  (i.e.,  $w^{t_{i+1}} = w^{t_i + \sigma_i}$ ), can be calculated as

$$q_j = \#\mathcal{I}_j^* = \# \left\{ \bigcup_{i=1}^j s^{t_i} \cap \{s^{t_1} + \sum_{i=1}^j \sigma_i\} \right\}, \tag{6}$$

where all steps  $\sigma_i$  are fixed for  $i = 1, \dots, j$  and  $j = 1, \dots, c^* - 1$ . Note that the sets  $\mathcal{I}_{j-1}^*$  ( $j \geq 1$ ) correspond to outputs  $w^{t_j}$  (where  $\mathcal{I}_0^* = \mathcal{I}_0 \stackrel{(4)}{=} s^{t_1}$ ). The sampling instances are given as  $t_j = t_1 + \sum_{i=1}^{j-1} \sigma_i$ , or  $t_j = t_{j-1} + \sigma_{j-1}$ , where  $t_{j-1} = t_1 + \sum_{i=1}^{j-2} \sigma_i$  is fixed. Similarly to the *GFSGA* with a constant sampling distance, the attack complexity is estimated as

$$T_{Comp.}^* = 2^{n-m} \times 2^{n-m-q_1} \times \dots \times 2^{n-m-q_{c^*-1}} \times L^3. \tag{7}$$

Remark 2 also applies here, thus if  $n-m-q_j \leq 0$  for some  $j \in \{1, \dots, c^*-1\}$ , then the knowledge of these  $q_j$  bits allows the attacker to uniquely identify the exact preimage value of the observed output, i.e., we have  $2^{(n-m-q_j)} = 1$  when  $n-m-q_j \leq 0$ . Notice that if the sampling steps  $\sigma_i$  are equal, i.e., they have a constant value  $\sigma = \sigma_i$ , for  $i = 1, 2, \dots, c^* - 1$ , we get  $q_i = r_i$ ,  $c^* = c$ ,  $R^* = R$  and  $T_{Comp.}^* = T_{Comp.}$ .

**Remark 4** It was already mentioned that the analysis of complexity  $T_{Comp.}^*$  appears to be very difficult, mainly due to the following reasons. For fixed  $m, n$  and  $L$ , it is clear that there exists a trade-off between the parameters  $q_j$  ( $j = 1, \dots, c^* - 1$ ) and  $c^*$ . More precisely, for larger values  $c^*$  we have that  $2^{n-m-q_j} > 1$  which implies the increase of  $T_{Comp.}^*$ , unless  $n-m-q_j \leq 0$ . For this reason, the optimal step(s) of the *GFSGA* attack (whether we consider a constant or variable mode of the attack) is the one which minimizes  $c^*$  satisfying at the same time the inequality  $nc^* - R^* > L$ . Furthermore, in the case of the constant *GFSGA* mode, the parameter  $c$  for which  $nc - R > L$  holds is not known prior to the completion of the sampling process. This also holds for the variable *GFSGA* mode, if we fix a sequence of sampling steps in advance. This, in combination with [22, Remark 3], give more insight how complicated the relation between parameters  $m, n, L, q_j, c^*$  and  $T_{Comp.}^*$  is.

### 3.1 The number of repeated equations for $GFSGA^*$

The relation between the number of repeated equations and complexity in the case of  $GFSGA$  has been analyzed in [22], where an alternative method for calculating the number of repeated equations has been derived. Similarly, in this section we derive an alternative method for calculating the number of repeated equations for  $GFSGA^*$  (Proposition 2).

For a given set of tap positions  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\}$ , let us consider the set of differences between the consecutive tap positions, i.e.,

$$D = \{d_j \mid d_j = l_{j+1} - l_j, j = 1, 2, \dots, n-1\}.$$

Based on this set the so-called scheme of all possible differences was defined in [22] as  $D^{\mathcal{I}_0} = \{l_j - l_k : l_j, l_k \in \mathcal{I}_0, l_j > l_k\}$  and used in [22] to calculate the number of repeated equations for  $GFSGA$ .

**Proposition 1** [22] *Let  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\}$  be a set of tap positions, and let*

$$D = \{l_{j+1} - l_j \mid j = 1, 2, \dots, n-1\} = \{d_1, d_2, \dots, d_{n-1}\}.$$

*The number of repeated equations is calculated as*

$$R = \sum_{i=1}^{n-1} \left( c - \frac{1}{\sigma} \sum_{k=i}^m d_k \right), \quad (8)$$

where  $\sigma \mid \sum_{k=i}^m d_k$  for some  $m \in \mathbb{N}$ ,  $i \leq m \leq n-1$  and  $\frac{1}{\sigma} \sum_{k=i}^m d_k \leq c-1$ . Moreover, if  $\frac{1}{\sigma} \sum_{k=i}^m d_k \geq c$ , for some  $1 \leq i \leq n-1$ , then  $(c - \frac{1}{\sigma} \sum_{k=i}^m d_k) = 0$ . This means that the repetition of the same equations (bits) starts to appear after the LFSR state  $s^{tc}$ .

In the case of  $GFSGA^*$ , the scheme of differences can also be used to calculate the number of repeated equations  $R^*$ . However, in this case the calculation is slightly more complicated compared to  $GFSGA$ , due to the fact that we have a variable step of sampling.

To illustrate the difference, let us consider the set of tap positions given by  $\mathcal{I}_0 = \{3, 5, 10, 14, 16\}$  ( $L = 20$  and  $n, m$  not specified). The corresponding set of consecutive differences is given as  $D = \{2, 5, 4, 2\}$ . The scheme of all differences related to  $D^{\mathcal{I}_0}$  is given as:

Table 1: The scheme of all differences for  $D = \{2, 5, 4, 2\}$ .

	Col. 1	Col. 2	Col. 3	Col. 4
$l_{j+1} - l_j$	2	5	4	2
$l_{j+2} - l_j$	7	9	6	
$l_{j+3} - l_j$	11	11		
$l_{j+4} - l_j$	13			

In addition, let us assume that the first two steps of sampling (distances between observed outputs) are given as:  $\sigma_1 = 5$ ,  $\sigma_2 = 2$ . To find the number of repeated bits (equations), we use the recursion of the sets  $\mathcal{I}_k^*$  given by relation (5). Even though  $c^*$  is the number of outputs for which an overdefined system can be set up, our purpose is to demonstrate the procedure of

finding repeated bits for  $\sigma_1, \sigma_2$ . The computation of the number of repeated bits is as follows:  
 1) The state bits at tap positions at time  $t_1$  correspond to  $\mathcal{I}_0 = \{l_1, l_2, l_3, l_4, l_5\} = \{3, 5, 10, 14, 16\}$ , thus  $s^{t_1} = (s_2, s_4, s_9, s_{13}, s_{15}) \stackrel{(4)}{=} \mathcal{I}_0$ . Since the first sampling distance  $\sigma_1 = 5$ , we consider the LFSR state  $s^{t_2} = \{\mathcal{I}_0 + \sigma_1\}$  which is given as

$$s^{t_2} = \{\mathcal{I}_0 + 5\} = (s_7, s_9, s_{14}, s_{18}, s_{20}) \stackrel{(4)}{=} \{8, 10, 15, 19, 21\}.$$

We obtain that  $\mathcal{I}_1^* = s^{t_1} \cap s^{t_2} = \mathcal{I}_0 \cap (\mathcal{I}_0 + 5) = \{10\}$ , which means that  $1 = \#\mathcal{I}_1^* = q_1$  bit is repeated and found in  $s^{t_2}$  from the first state  $s^{t_1} = \mathcal{I}_0$ .

In terms of the scheme of differences, this repetition corresponds to  $d_2 = l_3 - l_2 = \sigma_1 = 5$ , found as the first entry in Col. 2. In addition, note that  $d_2 = 5$  is the only entry in the scheme of differences  $D^{\mathcal{I}_0}$  which is equal to  $\sigma_1$ . The main difference compared to GFSGA is that in this case we do NOT consider the divisibility by  $\sigma_1$  in the scheme of differences due to variable sampling steps.

2) For  $\sigma_2 = 2$ , the observed outputs  $w^{t_2}$  and  $w^{t_3}$  satisfy  $w^{t_3} = w^{t_2 + \sigma_2} = w^{t_1 + \sigma_1 + \sigma_2}$ . The LFSR state  $s^{t_3}$ , which corresponds to the output  $w^{t_3}$ , is given as  $s^{t_3} = \{s^{t_2} + \sigma_2\} = \{s^{t_1} + (\sigma_1 + \sigma_2)\}$ , and therefore

$$s^{t_3} = \{s^{t_2} + 2\} = (s_9, s_{11}, s_{16}, s_{20}, s_{22}) = \{10, 12, 17, 21, 23\}.$$

At this position we check whether there are repeated bits from the LFSR state  $s^{t_2}$ , but also from the state  $s^{t_1}$ . To find all bits which have been repeated from the state  $s^{t_2}$ , we consider the intersection  $s^{t_3} \cap s^{t_2} = \{10, 21\}$ . In addition, the bits which are repeated from the state  $s^{t_1}$  are given by the intersection  $s^{t_3} \cap s^{t_1} = \{10\}$ . Hence, we have the case that the same bit, indexed by 10, has been shifted from the state  $s^{t_1}$  (since  $\sigma_1 + \sigma_2 = 7$  so that  $3 + 7 = 10$ ) and from  $s^{t_2}$  (since  $\sigma_2 = 2$ ) to  $s^{t_3}$ . On the other hand, the intersection corresponding to 21 gives us an equation that has not been used previously. Thus, the number of known state bits used in the reduction of the preimage space is 2, and therefore  $|S_{w^{t_3}}| = 2^{n-m-2}$ . In terms of the scheme of differences, one may notice that we have  $d_1 = l_2 - l_1 = 2 = \sigma_2$  (which refers to repetition from  $s^{t_2}$  to  $s^{t_3}$ ) and which gives  $d_1 = 2$  in Col. 1 and Col. 4. On the other hand, for  $d_1 + d_2 = l_3 - l_1 = 7 = \sigma_1 + \sigma_2$  (which refers to repetition from  $s^{t_1}$  to  $s^{t_3}$ ) the repeated bit is found in the same column, namely Col. 1, and therefore it is not counted. In general, we conclude here that if we have a matching of  $\sigma^{(2)} = d_2$  and  $\sigma^{(1)} = d_1 + d_2$  with some numbers (entries) in the scheme of differences which are in the same column (as we have here  $d_1 = \sigma^{(2)}$  and  $d_1 + d_2 = \sigma^{(1)}$ ), we calculate only one repeated bit in total from this column. If there were more than 2 matchings, the same reasoning applies and thus we would calculate only one repeated bit.

The procedure above may continue for any number of observed outputs at any sampling distances. For instance, if we consider some sampling step  $\sigma_j$  ( $j \geq 1$ ), then we also need to consider all sums of the steps  $\sigma^{(j-i)} = \sum_{h=j-i}^j \sigma_h$ , for all  $i = 0, \dots, j-1$  ( $j \geq 1$ ) and their matchings with some entries in the scheme of differences. In general, every repeated bit means that some sum(s) of steps  $\sigma^{(j-i)} = \sum_{h=j-i}^j \sigma_h$ ,  $i = 0, \dots, j-1$  is equal to some difference  $\sum_{p=r}^m d_p$ , for some  $m \in \mathbb{N}$ , over different columns, where  $r = 1, \dots, n-1$  relate to the columns in the scheme of differences. A total number of repeated bits  $R^*$  is the sum of all repeated bits over observed outputs at distances  $\sigma_j$  ( $j \geq 1$ ). Note that the method for calculation of the number of repeated bits

described above actually generalizes Proposition 1, since the use of constant sampling distance is just a special case of variable sampling.

**Proposition 2** *Let  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\}$  be a set of tap positions, and let*

$$D = \{l_{i+1} - l_i \mid i = 1, 2, \dots, n-1\} = \{d_1, d_2, \dots, d_{n-1}\}.$$

*Denoting  $\sigma^{(j-i)} = \sum_{h=j-i}^j \sigma_h$ ,  $i = 0, \dots, j-1$ , the number of repeated equations is calculated as*

$$R^* = \sum_{j=1}^{c^*-1} q_j = \sum_{j=1}^{c^*-1} \left( \sum_{i=0}^{j-1} \frac{1}{\sigma^{(j-i)}} \sum_{p=r}^m d_p \right), \quad (9)$$

*where the term  $\frac{1}{\sigma^{(j-i)}} \sum_{p=r}^m d_p = 1$  if and only if  $\sigma^{(j-i)} = \sum_{p=r}^m d_p$  for some  $m, r \in \mathbb{N}$ ,  $1 \leq r \leq m \leq n-1$ , otherwise it equals 0. If for a fixed  $r$  and different numbers  $m$  we have more matchings  $\sigma^{(j-i)} = \sum_{p=r}^m d_p$ , then only one bit will be taken in calculation.*

Clearly, the numbers  $m$  and  $r$  depend on the values  $\sigma^{(j-i)}$ ,  $i = 0, \dots, j-1$  ( $j \geq 1$ ) since we only consider those numbers  $m, r \in \mathbb{N}$  such that  $\sum_{p=r}^m d_p$  is equal to  $\sigma^{(j-i)}$ .

### 3.2 Two specific modes of $GFSGA^*$

In this section we present two modes of  $GFSGA^*$ , which in comparison to  $GFSGA$  depend less on the choice of tap positions. First we start with a general discussion regarding the attack complexity.

In order to obtain a minimal complexity of the  $GFSGA^*$  attack, it turns out that the main problem is actually a selection of the cipher outputs. This problem is clearly equivalent to the problem of selecting the steps  $\sigma_i$  which gives the minimal complexity  $T_{Comp}^*$ . The number of repeated bits (equations) at time instance  $t_i$  (for some  $i > 1$ ) always depends on all previous sampling points at  $t_1, \dots, t_{i-1}$ . This property directly follows from (5), i.e., from the fact that we always check the repeated bits which come from the tap positions of LFSR at time instances  $t_1, \dots, t_{i-1}$ . This means that the number of repeated bits  $q_i$  at time instances  $t_i$ , given by (6), always depends on the previously chosen steps  $\sigma_1, \dots, \sigma_{i-1}$ . This also implies that we cannot immediately calculate the number of required keystream blocks  $c^*$  for which the inequality  $nc^* - R^* > L$  is satisfied. This inequality can only be verified subsequently, once the sampling distances and the number of outputs  $c^*$  have been specified. Therefore we pose the following problem.

**Open Problem 1** *For a given set of tap positions  $\mathcal{I}_0 = \{l_1, \dots, l_n\}$ , without the knowledge of  $c^*$ , determine an optimal sequence of sampling distances  $\sigma_i$  for which the minimal complexity  $T_{Comp}^*$  is achieved.*

In what follows we provide two modes of the  $GFSGA^*$  whose performance will be discussed in Section 4.

### 3.3 $GFSGA_{(1)}^*$ mode of attack

In order to minimize the complexity  $T_{Comp}^*$ , one possibility is to maximize the values  $q_j$ , given by (6), by choosing suitable  $\sigma_i$ . However, this approach implies a trade-off between the values  $q_j$  and  $c^*$ , since  $c^*$  is not necessarily minimized. More precisely:

1) For the first step  $1 \leq \sigma_1 \leq L$  we take a value for which  $q_1$  is maximized, i.e., for which the cardinality

$$q_1 = \#\{s^{t_1} \cap (s^{t_1} + \sigma_1)\} = \#\{s^{t_1} \cap s^{t_2}\}$$

is maximized. Without loss of generality, we can take the minimal  $\sigma_1$  for which this holds.

2) In the same way, in the second step we take a value  $1 \leq \sigma_2 \leq L$  for which

$$q_2 = \#\{(s^{t_1} \cup s^{t_2}) \cap s^{t_3}\} = \#\{(s^{t_1} \cup s^{t_2}) \cap ((s^{t_1} + \sigma_1) + \sigma_2)\}$$

is maximized. As we know, the step  $\sigma_1$  here is fixed by the previous step. We continue this procedure until an overdefined system is obtained.

In other words, the values  $q_j$  are determined by the *maximum* function over  $\sigma_i$ , for  $1 \leq \sigma_i \leq L$ , i.e., we choose the steps  $\sigma_i$  for which we have:

$$q_j = \max_{1 \leq \sigma_j \leq L} \# \left\{ \bigcup_{i=1}^j s^{t_i} \cap \{(s^{t_1} + \sum_{i=1}^{j-1} \sigma_i) + \sigma_j\} \right\}, \quad (10)$$

where  $\sum_{i=1}^{j-1} \sigma_i$  is fixed and  $j = 1, \dots, c^* - 1$ . Hence, the function  $\max_{1 \leq \sigma_j \leq L}$  used in (10) means that we are choosing  $\sigma_i$  so that the maximal intersection of  $s^{t_{j+1}}$  with all previous LFSR states  $s^{t_1}, \dots, s^{t_j}$  (in terms of cardinality) is achieved. This mode of  $GFSGA^*$  we denote by  $GFSGA_{(1)}^*$ .

### 3.4 $GFSGA_{(2)}^*$ mode of attack

Another mode of  $GFSGA^*$ , based on the use of sampling distances that correspond to the differences between consecutive tap positions, is discussed in this section. The selection of steps  $\sigma_i$  and the calculation of repeated bits is performed as follows..

For a given set of tap positions  $\mathcal{I}_0 = \{l_1, \dots, l_n\}$ , let  $D = \{d_1, \dots, d_{n-1}\}$  be the corresponding set of differences between the consecutive tap positions. The sequence of sampling distances  $\sigma_i$  between the observed outputs  $w^{t_i}$  and  $w^{t_{i+1}}$  is defined as:

$$\begin{cases} \sigma_{1+p(n-1)} = d_1, \\ \sigma_{2+p(n-1)} = d_2, \\ \vdots \\ \sigma_{n-1+p(n-1)} = d_{n-1}, \end{cases} \quad (11)$$

for  $p = 0, 1, 2, \dots$ . That is, the first  $n - 1$  sampling distances are taking values exactly from the set  $D$  so that  $\sigma_1 = d_1, \sigma_2 = d_2, \dots, \sigma_{n-1} = d_{n-1}$ . Then, the next  $n - 1$  sampling distances are again  $\sigma_n = d_1, \sigma_{n+1} = d_2, \dots, \sigma_{2n-2} = d_{n-1}$ , and so on. This mode of the  $GFSGA^*$  we denote as  $GFSGA_{(2)}^*$ . For this mode, using Proposition 2, we are able to calculate a lower bound on the

number of repeated equations for every sampling step. Recall that at some sampling instance  $t_j$  there are some repeated bit(s) if and only if  $\sigma^{(j-i)} = \sum_{h=j-i}^j \sigma_h$ ,  $i = 0, 1, \dots, j-1$ , is equal to some  $\sum_{p=r}^m d_p = l_m - l_r$ , for some  $1 \leq r \leq m \leq n-1$ . In addition, if in the same column in the scheme of differences (which is equivalent to considering a fixed  $r$  and all the values  $m \geq r$ ) we have more matchings  $\sigma^{(j-i)} = \sum_{p=r}^m d_p$ , then only one bit is counted.

Hence, taking the first  $n-1$  sampling steps to be  $\sigma_1 = d_1, \sigma_2 = d_2, \dots, \sigma_{n-1} = d_{n-1}$ , the scheme of differences (constructed for our set  $D = \{d_1, \dots, d_{n-1}\}$ ) and Proposition 2 imply that  $q_1 \geq 1$ , since at least  $\sigma_1 = d_1$  is in Col. 1. Then  $q_2 \geq 2$ , since we have  $\sigma^{(2-0)} = \sigma^{(2)} = \sigma_2$  is equal to  $d_2$  in Col. 2., and  $\sigma^{(2-1)} = \sigma^{(1)} = \sigma_1 + \sigma_2$  is equal to  $d_1 + d_2$  in Col. 1. Continuing this process we obtain  $q_3 \geq 3, \dots, q_{n-1} \geq n-1$ , which in total gives at least

$$1 + 2 + \dots + (n-1) = \frac{n(n-1)}{2}$$

repeated equations for the first  $n-1$  observed outputs. In the same way, at least  $\frac{n(n-1)}{2}$  of bits are always repeated if further sampling at  $n-1$  time instances is performed in accordance to (11). For instance, when  $p = 1$  in (11) we have  $\sigma_n = d_1$ . In general, for  $1 \leq i \leq n-1$  and  $p \geq 0$  we have  $q_{i+p(n-1)} \geq i$ , where the sampling steps are defined by (11). We conclude this section with the following remarks.

**Remark 5** Both  $GFSGA_{(1)}^*$  and  $GFSGA_{(2)}^*$  depend less on the placement of the tap positions in comparison to  $GFSGA$ . Indeed, for both modes the sampling distances  $\sigma_i$  are selected with respect to a given placement of tap positions but regardless of what this placement in general might be. These modes are therefore more useful for cryptanalytic purposes rather than to be used in the design of an optimal allocation of tap positions (given the length of LFSR and the number of taps).

**Remark 6** In the case of  $GFSGA$  where we have equal distances between the observed outputs, one may notice that the sequence of numbers  $r_i = \#\mathcal{I}_i$  is always an increasing sequence. On the other hand, the sequence of numbers  $q_i = \#\mathcal{I}_i^*$  for  $GFSGA^*$  may not be increasing at all. In connection to Open problem 1, neither  $GFSGA_{(1)}^*$  nor  $GFSGA_{(2)}^*$  automatically provides an optimal sequence of steps  $\sigma_i$  (which would imply the minimization of  $T_{Comp}^*$ ). This means that there exist cases in which any of the modes  $GFSGA$ ,  $GFSGA_{(1)}^*$  and  $GFSGA_{(2)}^*$  may outperform the other two (cf. Table 3 and Table 4, Section 4.2).

## 4 Comparision between $GFSGA$ , $GFSGA_{(1)}^*$ and $GFSGA_{(2)}^*$

In this section we compare the performance of the three  $GFSGA$  modes when the tap positions are selected (sub)optimally by using the algorithms originally proposed in [22]. Moreover, the case when the set of differences  $D$  forms a full positive difference set is also considered and compared to the algorithmic approach.

### 4.1 Overview of the algorithms for tap selection

As briefly mentioned in the introduction the concept of a full difference set, which ensures that all the entries in the set of all pairwise differences are different, is not a very useful criterion



for tap selection. This is especially true when GFSGA-like cryptanalysis is considered as shown in Section 4.2. The same applies to the set of consecutive differences which may be taken to have mutually coprime entries which still does not ensure a sufficient cryptographic strength. Thus, there is a need for a more sophisticated algorithmic approach for designing (sub)optimally the placement of  $n$  tap positions for a given length  $L$  of the LFSR. The main idea behind the algorithms originally proposed in [22], presented below for self-completeness, is the use of the standard *GFSGA* mode with a constant sampling rate for the purpose of finding (sub)optimal placement of tap positions.

The proposed algorithms for the selection of tap positions use the design rationales that maximize the resistance of the cipher to the standard mode of *GFSGA*. Instead of specifying the set  $\mathcal{I}_0$ , both methods aim at constructing the set  $D$  of consecutive differences which gives a low number of repeated equations (confirmed by computer simulations) for any constant sampling distance  $\sigma$ , which implies a good resistance to GFSGA-like methods. As an illustration of the above mentioned algorithms several examples were provided in [22] and in particular an application to the stream cipher SOBER-t32 [11] and SFINKS [27] were analyzed in details. In what follows, we briefly recall the algorithms from [22] and then we discuss their structural properties related to the presented modes of GFSGA as well as to the criteria for tap selection proposed in [9] which provides a resistance to inversion attacks.

In both algorithms a quality measure for the choice of tap positions is the request that the optimal step (see Remark 3) of the *GFSGA* attack is as small as possible. The selection of tap positions itself is governed by the general rule, which is achieving co-prime differences between the tap positions (**Step A** in [22]) together with distributing taps all over the register (thus maximizing  $\sum_{i=1}^{n-1} d_i \leq L-1$ , where  $d_i = l_{i+1} - l_i$ ,  $\mathcal{I}_0 = \{l_1, \dots, l_n\}$  being a set of tap positions).

The first algorithm is designed to deal with situations when the size  $D$  is not large (say  $\#D \leq 10$ ). In this case the algorithm performs an exhaustive search of all permutations of the set  $D$  (**Step B** in [22]) and gives as an output a permutation which ensures a maximal resistance to *GFSGA*. The complexity of this search is estimated as  $O(n! \cdot K)$ , where  $K$  corresponds to the complexity of calculation  $T_{Comp.}$  for all possible  $\sigma$ .

**Remark 7** As mentioned in [22], to measure the quality of a chosen set of differences  $D$  with respect to the maximization of  $T_{Comp.}$  over all  $\sigma$ , the computer simulations indicate that an optimal ordering of the set  $D$  implies a small value of an optimal sampling distance  $\sigma$ . When choosing an output permutation (cf. Step 5 below), we always consider both  $\sigma$  and  $T_{Comp.}$  though  $\sigma$  turns out to be a more stable indicator of the quality of a chosen set  $D$ .

For practical values of  $L$ , usually taken to be  $L = 256$ , the time complexity of the above algorithm becomes practically infeasible already for  $n > 10$ . Using the previous algorithm, the second algorithm proposed in [22] regards a case when  $\#D$  is large. To reduce its factorial time complexity, we use the above algorithm only to process separately the subsets of the multiset  $D$  within the feasibility constraints imposed on the cardinalities of these subsets. The steps of this modified algorithm are given as follows.

**STEP 1:** Choose a set  $X$  by **Step A**, where  $\#X < \#D$  for which **Step B** is feasible;

**STEP 2:** Find the best ordering of  $X$  using the algorithm in **Step B** for

$$L_X = 1 + \sum_{x_i \in X} x_i < L \text{ and } m_X = \lfloor \#X \cdot \frac{m}{n-1} \rfloor;$$

**STEP 3:** Choose a set  $Y$  by **Step A**, where  $\#Y < \#D$  for which **Step B** is feasible;

- STEP 4:** “Generate” a list of all permutations of the elements in  $Y$ ;
- STEP 5:** Find a permutation ( $Y_p$ ) from the above list such that for a fixed set  $X$ , the new set  $Y_pX$  obtained by joining  $X$  to  $Y_p$ , denoted by  $Y_pX$  (with the parameters  $L_{Y_pX} = 1 + \sum_{x_i \in X} x_i + \sum_{y_i \in Y_p} y_i \leq L$  and  $m_{Y_pX} = \lfloor \#Y_pX \cdot \frac{m}{n-1} \rfloor$ ), allows a small optimal step  $\sigma$ , in the sense of Remark 7;
- STEP 6:** If such a permutation, resulting in a small value of  $\sigma$ , does not exist in Step 5, then back to Step 3 and choose another set  $Y$ ;
- STEP 7:** Update the set  $X \leftarrow Y_pX$ , and repeat the steps 3 - 5 by adjoining new sets  $Y_p$  until  $\#Y_pX = n - 1$ ;
- STEP 8:** Return the set  $D = Y_pX$ .

**Remark 8** The parameters  $L_X$  and  $m_X$  are derived by computer simulations, where  $L_X$  essentially constrains the set  $X$  and  $m_X$  keeps the proportionality between the numbers  $m$ ,  $\#X$  and  $\#D = n - 1$ .

The main question which arises here is whether the performance of the mentioned algorithms above can be improved by using some of the new presented modes in the previous section. Unfortunately, the main reasons why  $GFSGA$  can not be replaced by  $GFSGA_{(1)}^*$  or  $GFSGA_{(2)}^*$  are the following:

- a. Apart from Remark 5, due to their low dependency on the choice of tap positions, neither  $GFSGA_{(1)}^*$  nor  $GFSGA_{(2)}^*$  mode (through Proposition 2) simply do not provide enough information that can be used to construct the tap positions with high resistance to GFSGA attacks in general. It is clear that the presented algorithms above only give a (sub)optimal placement of tap positions due to impossibility to test exhaustively all permutations of  $D$  and additionally to perform testing of all difference sets  $D$  is infeasible as well. An optimal placement of tap positions providing the maximum resistance to GFSGA attacks leads us back to Open problem 1.
- b. One may notice that the main role of the constant step  $\sigma$  used in the design of the algorithm in [22] is to reduce the repetition of bits in general, since  $\sigma$  may take any value from 1 to  $L$ . This reduction of the repeated bits is significantly larger when using the constant step than any variable step of sampling in  $GFSGA_{(1)}^*$  or  $GFSGA_{(2)}^*$ , due to their specific definitions given by (10) and (11).

Notice that some criteria for tap selection regarding the resistance to the inversion attacks were proposed in [9]. The difference between the first and last tap position should be near or equal to  $L - 1$ , which turns out to be an equivalent criterion of maximization of the sum  $\sum_{i=1}^{n-1} d_i \leq L - 1$ , as mentioned above. Generalized inversion attacks [10] performed on filter generators, with the difference between the first and last tap position equal to  $M (= l_n - l_1)$ , have the complexity approximately  $2^M$  [10]. Hence, taking that  $\sum_{i=1}^{n-1} d_i = L - 1$ , where  $d_i = l_{i+1} - l_i$  (with tap positions  $\mathcal{I}_0 = \{l_1, \dots, l_n\}$ ), one of the criteria which thwarts (generalized) inversion attacks is easily satisfied.

In addition, one may also use a  $\lambda$ th-order full positive difference set [9] for tap selection, that is, the set of tap positions  $\mathcal{I}_0 = \{l_1, \dots, l_n\}$  with as small as possible parameter  $\lambda = \max_{1 \leq \sigma \leq M} |\mathcal{I}_0 \cap (\mathcal{I}_0 + \sigma)|$ . If  $\lambda = 1$ , then  $\mathcal{I}_0$  is a standard full positive difference set. As illustrated

in Table 4, to provide a high resistance to GFSGA-like attacks, the set of tap positions may be a  $\lambda$ th-order full positive difference set with higher values of  $\lambda$ . Note that in the case of inversion attacks, smaller  $\lambda$  is required. In other words, ( $\lambda$ th-order) full positive difference sets do not provide the same resistance to inversion attacks and GFSGA-like attacks, when the selection of tap positions is considered.

## 4.2 Full positive difference sets versus algorithms in [22]

In this section, we compare the performance of the three GFSGA modes by applying these attacks to a cipher whose tap positions are chosen using the algorithms in [22] and in the case the tap positions form suitably chosen full positive difference sets, respectively. We analyze the resistance to GFSGA attacks (using these two methods for tap selection) and conclude that full positive difference sets do not give an optimal placement of tap positions, i.e., they do not provide a maximal resistance to GFSGA-like cryptanalysis.

It is not difficult to see that the set of rules valid for the algorithms in [22] essentially require that we choose a set of tap positions  $\mathcal{I}_0$  so that the corresponding set of consecutive differences  $D$ , apart from having different elements (possibly all), is also characterized by the property that these differences are coprime and in a specific order. The situation when taps are not chosen optimally, implying a high divisibility of the elements in  $D$ , is illustrated in Table 2. Denoting by  $T_{Comp.}$ ,  $T_{Comp.(1)}^*$  and  $T_{Comp.(2)}^*$  the running time of the *GFSGA*, *GFSGA*<sub>(1)</sub><sup>\*</sup> and *GFSGA*<sub>(2)</sub><sup>\*</sup> mode, respectively, it is obvious that GFSGA is superior to other modes in most of the cases, as indicated in Table 2. In Table 3, we compare the performance of the three modes, if the tap

Table 2: Complexity comparison of all three GFSGA modes for "bad" tap choices.

$L$	$(n, m)$	$D$	$T_{Comp.}$	$T_{Comp.(1)}^*$	$T_{Comp.(2)}^*$
80	(9,2)	{12, 3, 6, 12, 6, 4, 24, 12}	$2^{43.97}$	$2^{67.97}$	$2^{62.97}$
120	(11,3)	{5, 10, 15, 4, 5, 10, 5, 15, 20, 25}	$2^{37.7}$	$2^{63}$	$2^{69.7}$
160	(15,6)	{14, 7, 3, 14, 7, 7, 14, 7, 14, 28, 7, 14, 14, 7}	$2^{32.97}$	$2^{32.97}$	$2^{50.97}$

positions (sets of differences  $D$ ) are chosen suboptimally according to rules and algorithms in [22]. Thus, if the tap positions are chosen according to the rules and algorithms in [22], it turns

Table 3: Complexity comparison of GFSGA modes - algorithmic selection of taps.

$L$	$(n, m)$	$D$	$T_{Comp.}$	$T_{Comp.(1)}^*$	$T_{Comp.(2)}^*$
80	(7,2)	{5, 13, 7, 26, 11, 17}	$2^{69.97}$	$2^{63.97}$	$2^{59.97}$
120	(13,3)	{5, 7, 3, 13, 6, 11, 5, 11, 7, 13, 21, 17}	$2^{99.7}$	$2^{104}$	$2^{78.7}$
160	(17,6)	{5,11,4,3,7,9,1,2,23,15,5, 13, 7, 26, 11, 17}	$2^{86.97}$	$2^{79.97}$	$2^{41.97}$
200	(21,7)	{3, 7, 9, 13, 18, 7, 9, 1, 2, 9, 1, 2, 23, 15, 5, 13, 7, 26, 11, 17}	$2^{108.9}$	$2^{96.93}$	$2^{68.93}$

out that *GFSGA*<sub>(1)</sub><sup>\*</sup> and *GFSGA*<sub>(2)</sub><sup>\*</sup> modes are more efficient than *GFSGA*.

In Table 4, we compare the resistance of a nonlinear filter generator (specified by  $L$ ,  $n$  and  $m$ ) to different GFSGA modes regarding the design rationales behind the choice of tap positions. Namely, for the same cipher (in terms of the parameters above), the attack complexities are evaluated for tap positions that form (suitable) full positive differences sets and, respectively, for the choices of tap positions given in Table 3 (with a slight modification adopted for different parameters  $L, n$  and  $m$ ). In general, the algorithmic approach gives a higher resistance to GFSGA-like cryptanalysis.

Table 4: Complexity comparison - full positive difference sets versus algorithmic choice.

$L$	$(n, m)$	Tap positions - Full positive difference sets		$T_{Comp.}$	$T_{Comp.(1)}^*$	$T_{Comp.(2)}^*$
80	(7,2)	{1, 3, 8, 14, 22, 23, 26}		$2^{35.97}$	$2^{37.97}$	$2^{57.97}$
120	(13,3)	{1, 3, 6, 26, 38, 44, 60, 71, 86, 90, 99, 100, 107}		$2^{86.72}$	$2^{90.72}$	$2^{95.72}$
160	(15,4)	{1, 5, 21, 31, 58, 60, 63, 77, 101, 112, 124, 137, 145, 146, 152}		$2^{96.97}$	$2^{105.97}$	$2^{116.97}$
200	(17,5)	{1, 6, 8, 18, 53, 57, 68, 81, 82, 101, 123, 139, 160, 166, 169, 192, 200}		$2^{113.93}$	$2^{123.93}$	$2^{132.93}$
$L$	$(n, m)$	Set of consecutive differences $D$ -algorithmic choice	$\lambda$	$T_{Comp.}$	$T_{Comp.(1)}^*$	$T_{Comp.(2)}^*$
80	(7,2)	{5, 13, 7, 26, 11, 17}	1	$2^{69.97}$	$2^{63.97}$	$2^{59.97}$
120	(13,3)	{5, 7, 3, 13, 6, 11, 5, 11, 7, 13, 21, 17}	3	$2^{99.7}$	$2^{104}$	$2^{78.7}$
160	(15,4)	{5, 3, 7, 1, 9, 17, 15, 23, 5, 13, 7, 26, 11, 17}	3	$2^{114.97}$	$2^{124.97}$	$2^{101.97}$
200	(17,5)	{7, 13, 10, 13, 7, 1, 9, 17, 15, 23, 5, 13, 7, 26, 11, 17}	3	$2^{120.93}$	$2^{120.93}$	$2^{113.93}$

**Remark 9** Table 4 also indicates that an algorithmic choice of tap positions may provide significantly better resistance against GFSGA-like attacks compared to full positive difference sets (for various parameters  $n, m$  and  $L$ ).

### 4.3 Further examples and comparisons

In this section we provide a few examples which illustrate the sampling procedure and specification of repeated bits for the  $GFSGA_{(1)}^*$  and  $GFSGA_{(2)}^*$  modes. In both cases we consider the set of consecutive differences  $D = \{5, 13, 7, 26, 11, 17\}$  (most of the differences being prime numbers) which corresponds to the set of tap positions  $\mathcal{I}_0 = \{1, 6, 19, 26, 52, 63, 80\}$ . We first consider the  $GFSGA_{(1)}^*$  mode.

**Example 1** Let the set of tap positions be given by  $\mathcal{I}_0 = s^{t_1} = \{1, 6, 19, 26, 52, 63, 80\}$ , where  $L = 80$  and  $F : GF(2)^7 \rightarrow GF(2)^2$  ( $n = 7, m = 2$ ). The set  $\mathcal{I}_0$  is chosen according to the algorithms in [22] and it is most likely an optimal choice of tap positions for the given parameters  $L, n$  and  $m$ . Recall that the variable sampling steps  $\sigma_i$  for  $GFSGA_{(1)}^*$  are determined by the maximum function used in relation (10). In Table 5, using the relation (10) we identify the repeated state bits until the inequality  $nc^* > L + R^*$  is satisfied for some  $c^*$ . The total number of repeated equations over all observed outputs is  $R^* = \sum_{k=1}^{c^*-1} q_k = 67$ , where the number of outputs is  $c^* = 22$ . Note that the first observed output  $w^{t_1}$  has the preimage space of full size,

$i$	Sets $\mathcal{I}_i^*$ (where $(k+1) \leftrightarrow s_k$ )	$q_i$	$\sigma_i$
1	{6}	1	5
2	{19, 24}	2	13
3	{26, 31, 44}	3	7
4	{52, 57, 70, 77}	4	26
5	{63, 68, 81, 88, 114}	5	11
6	{80, 85, 98, 105, 131, 142}	6	17
7	{85, 103}	2	5
8	{114, 147}	2	11
9	{131, 164, 175}	3	17
10	{118, 136}	2	5
11	{125, 138}	2	2
12	{131, 136, 182}	3	11
13	{138, 143, 156}	3	7
14	{164, 169, 182, 189}	4	26
15	{175, 180, 193, 200, 226}	5	11
16	{192, 197, 210, 217, 243, 254}	6	17
17	{197, 215}	2	5
18	{199, 217}	2	2
19	{210, 215, 261}	3	11
20	{217, 222, 235}	3	7
21	{243, 248, 261, 268}	4	26

Table 5: Repeated bits attained by sampling steps  $\sigma_i$  defined by (10).

and thus there are no repeated bits. Since we chose  $s^{t_1} = \mathcal{I}_0$ , the positions of repeated bits at the corresponding tap positions can be found and calculated as follows:

- The step  $\sigma_1 = 5$  gives the maximal intersection between  $s^{t_1}$  and  $s^{t_2} = \{s^{t_1} + 5\} = \{6, 11, 24, 31, 57, 68, 85\}$ , i.e., we have

$$\max_{1 \leq \sigma_1 \leq 80} \#\{s^{t_1} \cap (s^{t_1} + \sigma_1)\} = \max_{1 \leq \sigma_1 \leq 80} \#\{s^{t_1} \cap s^{t_2}\} = \{6\},$$

which yields  $q_1 = 1$ . The size of the preimage space is  $|S_{w^{t_2}}| = 2^{n-m-q_1} = 2^4$ .

- Assuming the knowledge of  $x^{t_2} \in S_{w^{t_2}}$  and  $x^{t_1} \in S_{w^{t_1}}$ , we search for an optimal shift  $\sigma_2$  of  $s^{t_2}$  so that  $q_2 = \#\{\{s^{t_1} \cup s^{t_2}\} \cap \{s^{t_2} + \sigma_2\}\}$  is maximized. Note that at this point,  $\{s^{t_1} + \sigma_1\} = s^{t_2}$  is fixed. This gives  $\sigma_2 = 13$  and  $q_2 = 2$ . The set of repeated bits is

$$\max_{1 \leq \sigma_2 \leq 80} \#\{\{s^{t_1} \cup s^{t_2}\} \cap \{s^{t_2} + \sigma_2\}\} = \{19, 24\},$$

since  $s^{t_3} = \{s^{t_2} + \sigma_2\} = \{19, 24, 37, 44, 70, 81, 98\}$ . The preimage space has the cardinality  $|S_{w^{t_3}}| = 2^{n-m-q_2} = 2^3$ .

In this way, we can completely determine the preimage spaces and the positions of the repeated bits. Since for  $i \in \{5, 6, 15, 16\}$  we have  $q_i \geq n - m = 5$ , then  $2^{n-m-q_i} = 1$  (by convention). Once the other values  $q_j$  have been computed, for  $j \in \{1, 2, \dots, 21\} \setminus \{5, 6, 15, 16\}$ , the attack complexity can be estimated as

$$T_{Comp.(1)}^* = 2^{n-m} \times 2^{n-m-q_1} \times \dots \times 2^{n-m-q_{21}} \times L^3 \approx 2^{63.97}.$$

In the case of GFSGA, an optimal choice of the sampling distance (cf. Remark 3) is any  $\sigma \in \{1, 13, 37\}$ . Each of these sampling steps requires  $c = 16$  observed outputs and gives  $R = 24$  repeated equations, where the set of all repeated bits is given by

$$\{r_1, r_2, \dots, r_{15}\} = \{0, 0, 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4\}.$$

The values  $r_i = 0$ ,  $i = 1, 2, 3, 4$ , mean that the corresponding sets  $\mathcal{I}_i$  are empty. The attack complexity of GFSGA is then estimated as  $T_{Comp.} \approx 2^{69.97}$ .

**Example 2** Now, for the same function  $F$  and the set of tap positions  $\mathcal{I}_0 = s^{t_1}$  (or the set of differences  $D = \{5, 13, 7, 26, 11, 17\}$ ), we illustrate the  $GFSGA_{(2)}^*$  mode. In Table 6, we show all repeated bits for the sampling steps  $\sigma_i$  of the  $GFSGA_{(2)}^*$  mode, which are defined by relation (11). Recall that the steps  $\sigma_i$  in this case are defined so that every  $n - 1 = 6$  outputs are at distances  $d_i \in D$ . By formula (7), the complexity of  $GFSGA_{(2)}^*$  is estimated as  $T_{Comp.(2)}^* \approx 2^{59.97}$ ,

Table 6: Repeated bits attained by sampling steps  $\sigma_i$  defined by (11).

$i$	Sets $\mathcal{I}_i^*$ (where $(k+1) \leftrightarrow s_k$ )	$q_i$	$\sigma_i$
1	{6}	1	5
2	{19, 24}	2	13
3	{26, 31, 44}	3	7
4	{52, 57, 70, 77}	4	26
5	{63, 68, 81, 88, 114}	5	11
6	{80, 85, 98, 105, 131, 142}	6	17
7	{85, 103}	2	5
8	{98, 103}	2	13
9	{105, 110, 123}	3	7
10	{131, 136, 149, 156}	4	26
11	{142, 147, 160, 167, 193}	5	11
12	{159, 164, 177, 184, 210, 221}	6	17
13	{164, 182}	2	5
14	{177, 182}	2	13
15	{184, 189, 202}	3	7
16	{210, 215, 228, 235}	4	26
17	{221, 226, 239, 246, 272}	5	11
18	{238, 243, 256, 263, 289, 300}	6	17
19	{243, 261}	2	5
20	{256, 261}	2	13
21	{263, 268, 281}	3	7

and thus this mode outperforms both GFSGA and  $GFSGA_{(1)}^*$ . The total number of repeated equations in this case is given by  $R^* = 72$ , for  $c^* = 22$  observed outputs. Notice that both modes  $GFSGA_{(1)}^*$  and  $GFSGA_{(2)}^*$  required in total 22 outputs to construct an overdefined system of linear equations ( $nc^* > L + R^*$ ).

## 5 Employing GFSGA in other settings

The main limitation of GFSGA-like attacks is their large complexity when applied to standard filtering generators that only output a single bit each time the cipher is clocked. In addition, this generic method cannot be applied in a straightforward manner in the cryptanalysis of ciphers that use NFSRs. In this section, we discuss the possibility of improving the efficiency and/or



applicability of GFSGA with variable sampling step for the above mentioned scenarios. It will be demonstrated that GFSGA with variable sampling step may be employed in combination with other cryptanalytic methods to handle these situations as well.

### 5.1 GFSGA applied to single-output nonlinear filter generators ( $m = 1$ )

The time complexity of GFSGA with variable sampling step is given by (7), i.e.,

$$T_{Comp}^* = 2^{n-m} \times 2^{n-m-q_1} \times \dots \times 2^{n-m-q_{c^*-1}} \times L^3,$$

and clearly when  $m = 1$  the complexity becomes quickly larger than the time complexity of exhaustive search (for some common choices of the design parameters  $n$  and  $L$ ). Based on annihilators in fewer variables of a nonlinear filtering function  $f(x_1, \dots, x_n)$ , Jiao *et al.* proposed another variant of FSGA in [14]. The core idea of this attack is to reduce the size of the preimage space via annihilators in fewer variables  $g_1(x_{j_1}, \dots, x_{j_d})$  and  $g_2(x_{j_1}, \dots, x_{j_d})$  such that  $f(x_1, \dots, x_n)g_1(x_{j_1}, \dots, x_{j_d}) = 0$  and  $(f(x_1, \dots, x_n) \oplus 1)g_2(x_{j_1}, \dots, x_{j_d}) = 0$ , where  $\{j_1, \dots, j_d\} \subset \{1, \dots, n\}$ . It was shown that the time complexity of this attack is given by

$$T_{Comp}^\Delta = \|S_{g_1=0}\|^{c_1} \times \|S_{g_2=0}\|^{c_2} \times L^\omega,$$

where  $\|S_{g_i=0}\|$ , for  $i = 1, 2$ , is the size of preimage space of the annihilator  $g_i$  (restricted to the variables  $\{j_1, \dots, j_d\}$ ),  $c^* = \lceil \frac{L}{d} \rceil$  is the number of sampling steps,  $c^* = c_1 + c_2$  and  $\omega = \log_2 7 \approx 2.807$  is the exponent of Gaussian elimination. In [14], it was also shown that this variant of FSGA could be applied to single-output filter generators. For instance, letting  $L = 87$  and using a nonlinear Boolean functions  $f(x_1, \dots, x_6)$  as in “Example 2” in [14], it was demonstrated that the time complexity of this attack is only about  $2^{80}$  operations, whereas the time complexity of FSGA is about  $2^{87}$  operations in [14].

In a similar manner, the same approach leads to a reduction of time complexity when GFSGA with variable sampling step is considered. For instance, let the set of tap positions be given by  $\mathcal{I}_0 = s^{t_1} = \{1, 6, 19, 26, 52, 63\}$  corresponding to the inputs  $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ , respectively. Using the filtering function  $f : \mathbb{F}_2^6 \rightarrow \mathbb{F}_2$  of “Example 2” in [14], one can deduce that  $\|S_{g_1(x_2, x_4, x_6)=0}\| = \|S_{g_2(x_2, x_4, x_6)=0}\| = 5$ . Actually, we can only consider the tap positions  $\{6, 26, 63\}$  with full positive difference set  $\{20, 37\}$ . Moreover, let us use the variable sampling steps  $\sigma_i = 20$  and  $\sigma_{i+1} = 37$ , alternately. In such a case, the preimage space of annihilator can be further reduced to  $\|S_{g_1(x_2, x_4, x_6)=0}^*\| = \|S_{g_2(x_2, x_4, x_6)=0}^*\| \approx 2.5$  by using the repeated bits. The time complexity of GFSGA with variable sampling steps is about  $5 \times 2.5^{42} \times 87^{2.807} \approx 2^{76.32} < 2^{80}$  operations. In particular, the number of variable sampling points is 43 since at the first sampling point 3 linear relations are obtained and the remaining 42 sampling points give  $2 \times 42 = 84$  linear relations, thus in total  $3 + 84 = 87 = L$  linear equations are derived. It directly means that our GFSGA with variable sampling step outperforms the variant of FSGA in [14].

Due to the small sized parameters  $L$  and  $n$  the above example does not illustrate a full potential of using GFSGA in cryptanalysis of single-output filtering generators. Its purpose is rather to show that GFSGA and its variants can be efficiently combined with other cryptanalytic methods. The most promising approach seems to be an interaction of GFSGA with algebraic attacks using small degree annihilators or restrictions of the filtering function  $f$ . Indeed, the use of repeated bits not only reduces the preimage space it essentially also fixes a subset of input

variables and therefore these restrictions of  $f$  may have annihilators of very low degree. This implies the existence of additional low degree equations in state bits which may be either used for checking the consistency of the linear system and after all (for sufficiently large number of fixed variables) these equations become linear. It is beyond the scope of this paper to investigate further the performance of this combined method but we believe that this kind of attack may become efficient against single-output filter generators with standard choice of the parameters  $L$  and  $n$ .

## 5.2 Applying GFSGA to NFSR-based ciphers

A current tendency in the design of stream ciphers, motivated by efficient hardware implementation, is the use of NFSRs in combination with (rather simple) nonlinear filtering function. For instance, this idea was employed in the design of the famous stream ciphers Trivium [7] and Grain family [2]. Apparently, none of the GFSGA variants can be applied for recovering the initial state of these ciphers but rather for deducing certain internal state of the cipher. In this scenario, the complexity of GFSGA is directly related to the complexity of solving an over-defined system of low degree equations rather than a system of linear equations. More precisely, assuming that the length of NFSR is  $L$  bits, the algebraic degree of its update function is  $r$ , and the filtering function  $F : GF(2)^n \rightarrow GF(2)^m$ , then the time complexity of GFSGA is given by

$$T_{Comp}^{**} = 2^{n-m} \times 2^{n-m-q_1} \times \dots \times 2^{n-m-q_{c^*-1}} \times D^\omega, \quad (12)$$

where  $D = \sum_{i=0}^{e \times r} \binom{L}{i}$ ,  $\omega = 2.807$  is the coefficient of Gaussian elimination,  $c^*$  is the number of sampling steps, and  $e$  is closely related to the parameters  $n, m, L, c^*$  and specified tap positions. The complexity being much larger than for LFSR-based ciphers, due to the term  $D^\omega$ , makes GFSGA methods practically inefficient.

However, one may mount another mode of internal state recovery attack which employs the GFSGA sampling procedure, but without solving systems of equations at all. More precisely, this new type of internal state recovery attack also employs the sampling of outputs within a certain sampling window which then allows us to efficiently recover a certain portion of internal state bits from the reduced preimage spaces corresponding to the observed outputs. To describe the attack in due detail, let us denote by  $p$  the distance between the last entry of NFSR (where NFSR is updated) and the tap position closest to this registry cell. We assume that this distance satisfies the inequality  $(p-1) \times n > L$ , where  $n$  is the number of inputs (tap positions) of a filtering function  $F : \mathbb{F}_2^n \rightarrow \mathbb{F}_m$ . Note that this condition implies that either  $p$  or  $n$  are relatively large. In such a case, let us choose the constant sampling steps  $\sigma_i = 1$ , for  $i = 1, \dots, p-1$ , and assume the adversary can directly recover, say  $R_p$  internal state bits (in total), at these  $p-1$  sampling instances. The remaining  $L - R_p$  internal state bits are still unknown and the adversary can exhaustively guess these bits to recover the whole internal state. The process of identifying the correct internal state is as follows. For each possible internal state candidate, a portion of  $L$  keystream bits  $Z^t = (z_1, \dots, z_L)$  at time instance  $t$  is determined using a given encryption algorithm. Then, comparing  $L$  keystream bits  $Z^{*t} = (z_1^*, \dots, z_L^*)$  at time instance  $t$  generated by the cipher (with unknown secret internal state), we can distinguish the correct and wrong internal states by directly checking if  $Z^t = Z^{*t}$ . In particular, if  $Z^t = Z^{*t}$ , then the guessed internal state would be the correct one, otherwise another internal state candidate is

considered. Consequently, the time complexity of this internal state recovery attack, assuming that remaining  $L - R_p$  bits are guessed, is given by

$$T_{Comp}^{**} = 2^{n-m} \times 2^{n-m-q_1} \times \dots \times 2^{n-m-q_{p-2}} \times 2^{L-R_p}. \quad (13)$$

The memory complexity of this attack is only  $(p-1) \times n \times 2^{n-1} + L$  bits, which are used to save all the element of preimage spaces and  $L$  keystream bits.

The following example illustrates an application of this approach to an NFSR-based cipher that largely resembles the NFSR used in the Grain-128 cipher. In particular, the process of recovering  $R_p$  internal state bits is described more thoroughly.

**Example 3** Let  $L = 128$ ,  $n = 8$ ,  $m = 1$ . The update function of NFSR is defined below (a slightly modified variant of the NFSR used in Grain-128 [2] without cubic and quartic terms):

$$\begin{aligned} b_{t+128} = & 1 \oplus b_t \oplus b_{t+26} \oplus b_{t+56} \oplus b_{t+91} \oplus b_{t+96} \oplus b_{t+3}b_{t+67} \oplus b_{t+11}b_{t+13} \\ & \oplus b_{t+17}b_{t+18} \oplus b_{t+27}b_{t+59} \oplus b_{t+40}b_{t+48} \oplus b_{t+61}b_{t+65} \oplus b_{t+68}b_{t+84}. \end{aligned}$$

The set of tap positions which we consider is given by  $\mathcal{I}_0 = s^{t_1} = \{l_1, \dots, l_8\} = \{1, 7, 21, 26, 52, 67, 89, 105\}$ , and it corresponds to a full positive difference set  $\{6, 14, 5, 26, 15, 22, 16\}$ . The distance between the last tap position and the NFSR update position is  $p = 128 - 105 = 23$ . This means that if we consider an updated internal state bit (the first nonlinear bit) and constant sampling rate  $\sigma_i = 1$ , this bit will appear at the tap position after 23 sampling instances. At the same time, employing the fact that many of these bits appear at some tap positions (thus using the idea of GFSGA), the adversary directly obtains many bits corresponding to 128-bit internal state as follows:

1. By relation (5) and sampling steps described in Section 3, collecting all repeated bits over  $p-1$  observed outputs (which are on consecutive distances  $\sigma_i = 1$ ), we determine all preimage spaces  $S_{w^{t_i}}$  ( $i = 1, \dots, p-1$ ) and their sizes.
2. Our approach implies that at the sampling instance  $i$  we recover (essentially guess)  $n - q_i$  internal state bits which must match to one of the  $2^{n-q_i-1}$  preimages. It is important to note that the bits which come from preimage spaces are the only candidates to be an internal state of the registry, since they are precisely determined by consecutive repetitions over  $p-1$  observed outputs.

Table 7 specifies the number of recovered internal state bits, and the sizes of corresponding preimage spaces. Denoting by  $R_p$  the total number of recovered bits, we can see (from Table 7) that  $R_p = 8 + 8 \times 4 + 7 + 6 \times 8 + 5 + 4 + 3 \times 6 = 122 < 128$ , where these bits are calculated using (5), for  $\sigma_i = 1$ , ( $i = 1, \dots, 22$ ).

The adversary can further guess the remaining  $L - R_p = 128 - 122 = 6$  internal state bits, and thus the time complexity, using (13), of this attack is about

$$T_{Comp}^{**} = 2^{7+7 \times 4+6+5 \times 8+4+3+2 \times 6} \times 2^6 = 2^{106} < 2^{128}.$$

The data complexity of this attack is only about  $22 + 128 = 150$  keystream bits. The memory complexity is upper bounded by  $22 \times 8 \times 2^7 + 128 < 2^{15}$  bits, which corresponds to storing at most  $2^7$  elements from preimage spaces and 128 keystream bits. The success rate is close to one since there are  $2^{7+7 \times 4+6+5 \times 8+4+3+2 \times 6+6} = 2^{106}$  internal state candidates in total and therefore only a small portion of about  $2^{106} \times 2^{-128} = 2^{-22} < 1$  wrong internal state candidates can pass the test.

Example 3 demonstrates that GFSGA-like attacks can be applied to NFSR-based stream ciphers without employing any structural properties of the filtering function. The following

Table 7: Recovered bits obtained by sampling step  $\sigma_i = 1$

$i$	Recovered bits of internal state	$q_i$	The size of preimage space
1	8	0	$2^7$
2	8	0	$2^7$
3	8	0	$2^7$
4	8	0	$2^7$
5	7	1	$2^6$
6	6	2	$2^5$
7	6	2	$2^5$
8	6	2	$2^5$
9	6	2	$2^5$
10	6	2	$2^5$
11	6	2	$2^5$
12	6	2	$2^5$
13	6	2	$2^5$
14	5	3	$2^4$
15	4	4	$2^3$
16	3	5	$2^2$
17	3	5	$2^2$
18	3	5	$2^2$
19	3	5	$2^2$
20	3	5	$2^2$
21	3	5	$2^2$

example illustrates an application of GFSGA to a hybrid NFSR/LFSR-based cipher whose design is very similar to Grain-128 cipher. The major difference is the key length, since our variant assumes that the key length is  $L = 256$  bits rather than 128-bit key used in Grain-128 [2].

**Example 4** Let  $L = 256$ ,  $n = 17$ ,  $m = 1$ . The internal state of our variant of Grain-128 consists of one 128-bit LFSR and one 128-bit NFSR, whose state bits are denoted by  $(s_0, \dots, s_{127})$  and  $(b_0, \dots, b_{127})$ , respectively. Their update functions are defined respectively as follows (see also [2]):

$$\begin{aligned}
s_{t+128} &= s_t \oplus s_{t+7} \oplus s_{t+38} \oplus s_{t+70} \oplus s_{t+81} \oplus s_{t+96} \\
b_{t+128} &= s_t \oplus b_t \oplus b_{t+26} \oplus b_{t+56} \oplus b_{t+91} \oplus b_{t+96} \oplus b_{t+3}b_{t+67} \oplus b_{t+11}b_{t+13} \\
&\quad \oplus b_{t+17}b_{t+18} \oplus b_{t+27}b_{t+59} \oplus b_{t+40}b_{t+48} \oplus b_{t+61}b_{t+65} \oplus b_{t+68}b_{t+84}
\end{aligned} \tag{14}$$

For this variant of Grain-128 cipher we consider the same set of tap positions that are used in the standard Grain-128 cipher, i.e., the tap position are  $A = \{2, 12, 15, 36, 45, 64, 73, 89, 95\}$  for the NFSR and  $B = \{8, 13, 20, 42, 60, 79, 93, 95\}$  for the LFSR. Note that the largest tap index in  $A$  is 95, and the NFSR is updated at position 128, i.e., their distance is  $p = 128 - 95 = 33$ . Similarly as in Example 3, sampling at the constant rate  $\sigma_i = 1$ , Table 8 specifies the number of recovered (repeated) bits of internal state, and the size of preimage spaces.

Thus, the adversary can directly obtain  $17 + 227 = 244 < 256$  internal state bits. The remaining  $L - R_p = 256 - 244 = 12$  internal state bits can then be guessed, which then leads to a recovery of the whole 256-bit internal state. Therefore, the time complexity of this attack is

Table 8: Repeated bits attained by sampling step  $\sigma_i = 1$

$i$	Recovered bits of internal state	$q_i$	The size of preimage space
1	17	0	$2^{16}$
2	16	1	$2^{15}$
3	15	2	$2^{14}$
4	15	2	$2^{14}$
5	14	3	$2^{13}$
6	13	4	$2^{12}$
7	12	5	$2^{11}$
8	12	5	$2^{11}$
9	10	7	$2^9$
10	9	8	$2^8$
11	9	8	$2^8$
12	9	8	$2^8$
13	9	8	$2^8$
14	8	9	$2^7$
15	8	9	$2^7$
16	7	10	$2^6$
17	7	10	$2^6$
18	6	11	$2^5$
19	4	13	$2^3$
20	4	13	$2^3$
21	3	14	$2^2$
22	2	15	2
23	2	15	2
24	2	15	2
25	2	15	2
26	2	15	2
27	2	15	2
28	2	15	2
29	2	15	2
30	2	15	2
31	2	15	2
$\Sigma = 227$			$\Pi = 2^{196}$

about

$$T_{Comp}^{**} = 2^{16+196} \times 2^{12} = 2^{224} < 2^{256}.$$

The above example demonstrates that GFSGA-like cryptanalysis is also applicable to hybrid NFSR/LFSR-based ciphers. In particular, it is shown that the tap positions have a very important impact on the security of NFSR/LFSR-based ciphers.

**Remark 10** *In difference to the time-memory-data trade-off attacks or algebraic attacks, this attack has more favorable data and memory complexity. For instance, in Example 4, the data complexity of this attack is only about  $32 + 229 = 261 \approx 2^8$  keystream bits. Namely, in the first step we use 32 sampling instances to determine all specified preimage spaces and their sizes under constant sampling rate  $\sigma_i = 1$ , and in the second step we need to use about 229 fresh keystream bits to determine the correct state. Notice that the memory complexity of this attack is only about  $32 \times 17 \times 2^{16} + 256 \approx 2^{25}$  bits. On the other hand, if the filtering function is  $f : GF(2)^{17} \rightarrow GF(2)^m, m > 4$ , then the time complexity of this attack is less than  $2^{128}$  operations. It implies that this attack would outperform the time-memory-data trade-off attack for  $m > 4$ .*

### 5.3 Grain-128 tap selection

We have already remarked that the tap selection for both SOBER-t32 and SFINKS was not optimal with respect to their resistance to GFSGA cryptanalysis, see also [22]. We show that the same is true when Grain-128 is considered, thus there exist better selections that ensure greater resistance to GFSGA-like cryptanalysis.

We assume that either LFSR or NFSR, whose tap positions are given in Example 4, of Grain-128 are used as state registers in a filter generator and we apply different modes of GFSGA to these schemes. In the case when the LFSR of Grain-128 is employed in such a scenario then the complexities of the three different modes of GFSGA are given as,

Table 9: Time complexity of different modes of GFSGA on LFSR of Grain-128

$T_{Comp.}$	$T_{Comp.(1)}^*$	$T_{Comp.(2)}^*$
$2^{108}$	$2^{125}$	$2^{118}$

Using our algorithm for finding a (sub)optimal placement of tap positions, instead of using the set  $A = \{2, 12, 15, 36, 45, 64, 73, 89, 95\}$ , we find another set of tap positions given as  $\{1, 16, 27, 54, 71, 95, 108, 127\}$  which gives the following complexities,

$$T_{Comp.} = 2^{129}, \quad T_{Comp.(1)}^* = 2^{132}, \quad T_{Comp.(2)}^* = 2^{123}.$$

A similar improvement can also be achieved when the tap positions of NFSR in Grain-128 are considered. In this case the original placement of tap positions (the set  $B$  in Example 4) gives the following complexities,

Table 10: Time complexity of different modes of GFSGA on NFSR of Grain-128

$T_{Comp.}$	$T_{Comp.(1)}^*$	$T_{Comp.(2)}^*$
$2^{114}$	$2^{125}$	$2^{122}$

On the other hand, our algorithm suggest somewhat better allocation of these taps given by  $\{3, 10, 29, 42, 59, 67, 88, 103, 126\}$ , which then induces the following complexities of the three GFSGA modes,

$$T_{Comp.} = 2^{130}, \quad T_{Comp.(1)}^* = 2^{139}, \quad T_{Comp.(2)}^* = 2^{125}.$$

## 6 Conclusions

In this article we have investigated the problem of selecting tap positions of the driving LFSR used in filter generators. The importance of this problem seems to be greatly neglected by the designers since to the best of our knowledge only some heuristic design rationales (such as the concepts of full positive difference sets and co-primality of consecutive differences) can be found in the literature. The algorithmic approach of selecting taps (sub)optimally given their number and the length of LFSR appears to generate good solutions for this problems, though its further



optimization and development may result in a better performance. Two additional modes of GFSGA have been introduced and it turns out that these modes in many cases can outperform the standard GFSGA mode. The idea of combining GFSGA technique and algebraic attacks appears to be a promising unified cryptanalytic method against LFSR-based stream ciphers though a more thorough analysis and some practical applications are needed for confirming its full potential.

## Acknowledgment

Samir Hodžić is supported in part by the Slovenian Research Agency (research program P3-0384 and Young Researchers Grant) and Enes Pasalic is partly supported by the Slovenian Research Agency (research program P3-0384 and research project J1-6720). Yongzhuang Wei was supported in part by the Natural Science Foundation of China (61572148), in part by the Guangxi Natural Science Found (2015GXNSFGA139007), in part by the project of Outstanding Young Teachers Training in Higher Education Institutions of Guangxi.

## References

- [1] R. ANDERSON. Searching for the optimum correlation attack. In *Fast Software Encryption, FSE 94*, vol. LNCS, pp. 137–143. Springer-Verlag, 1995.
- [2] M. AGREN, M. HELL, T. JOHANSSON, AND W. MEIER. A new version of Grain-128 with optional authentication. *International Journal of Wireless and Mobile Computing*, vol. 5, no. 1, pp. 48–59, 2011.
- [3] A. BIRYUKOV AND A. SHAMIR. Cryptanalytic time/memory/data tradeoffs for stream ciphers. In *Advances in Cryptology—ASIACRYPT 2000*, vol. LNCS 1976, pp. 1–13. Springer-Verlag, 2000.
- [4] A. BRAEKEN AND B. PRENEEL. Probabilistic algebraic attacks. In *IMA Conference on Cryptography and Coding*, vol. LNCS 3796, pp. 290–303. Springer-Verlag, 2005.
- [5] N. COURTOIS. Algebraic attacks on combiner with memory and several outputs. In *International Conference on Information Security and Cryptology – ICISC 2004*, vol. LNCS 3506, pp. 3–20. Springer-Verlag, 2005.
- [6] N. COURTOIS AND W. MEIER. Algebraic attacks on stream ciphers with linear feedback. In *Advances in Cryptology—EUROCRYPT 2003*, vol. LNCS 2656, pp. 346–359. Springer-Verlag, 2003.
- [7] C. D. CANNIÈRE AND B. PRENEEL. Trivium: A stream cipher construction inspired by block cipher design principles. In *Information Security*, vol. LNCS 4176, pp. 171–186. Springer-Berlin Heidelberg, 2006.
- [8] J. DJ. GOLIC. Intrinsic statistical weakness of keystream generators. In *Advances in Cryptology—ASIACRYPT 1994*, vol. LNCS 917, pp. 91–103. Springer-Verlag, 1995.

- [9] J. DJ. GOLIĆ. On the security of nonlinear filter generators. In *Fast Software Encryption '96*, vol. LNCS 1039, pp. 173–188. Springer-Verlag, 1996.
- [10] J. DJ. GOLIĆ, ANDREW CLARK, AND ED DAWSON. Generalized inversion attack on nonlinear filter generators. *IEEE Trans. Computers*, vol. 49, no. 10, pp. 1100–1109, 2000.
- [11] P. HAWKES AND G. ROSE. Primitive specification and supporting documentation for SOBER-t16 submission to NESSIE. In *Proceedings of the First Open NESSIE Workshop, KU-Leuven*, 2000.
- [12] M. HELLMAN. A cryptanalytic time-memory tradeoff. *IEEE Trans. on Inform. Theory*, vol. 26, no. 4, pp. 401–406, 1980.
- [13] J. HONG AND P SARKAR. New applications of time memory data tradeoffs. In *Advances in Cryptology—ASIACRYPT 2005*, vol. LNCS 3788, pp. 353–372. Springer-Verlag, 2005.
- [14] L. JIAO, M. WANG, Y. LI, AND M. LIU. On annihilators in fewer variables basic theory and applications. *Chinese Journal of Electronics*, vol. 22, no. 3, pp. 489–494, 2013.
- [15] J. L. MASSEY. Shift-register synthesis and BCH decoding. *IEEE Trans. on Inform. Theory*, vol. 15, no. 1, pp. 122–127, 1969.
- [16] W. MEIER AND O. STAFFELBACH. Fast correlation attacks on certain stream ciphers. In *J. of Cryptology*, vol. 1(3), pp. 159–176, 1989.
- [17] W. MEIER, E. PASALIC, AND C. CARLET. Algebraic attacks and decomposition of Boolean functions. In *Advances in Cryptology—EUROCRYPT 2004*, vol. LNCS 3027, pp. 474–491. Springer-Verlag, 2004.
- [18] A. J. MENEZES, P. C. VAN OORSCHOT, S. A. VANSTONE, R. L. RIVEST, Handbook of Applied Cryptography. 1997
- [19] M. J. MIHALJEVIĆ, M. P. C. FOSSORIER, AND H. IMAI. A general formulation of algebraic and fast correlation attacks based on dedicated sample decimation. In *Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, vol. LNCS 3857, pp. 203–212, Springer-Verlag, 2006.
- [20] M. J. MIHALJEVIĆ, S. GANGOPADHYAY, G. PAUL, AND H. IMAI. Internal state recovery of Grain-v1 employing normality order of the filter function In *IET Information Security*, vol. 6, no. 2, pp. 55–64, 2006.
- [21] K. NYBERG. On the construction of highly nonlinear permutations. In *Advances in Cryptology—EUROCRYPT'92*, vol. LNCS 658, pp. 92–98. Springer-Verlag, 1992.
- [22] E. PASALIC, S. HODŽIĆ, S. BAJRIĆ, Y. WEI. Optimizing the placement of tap positions. *Cryptography and Information Security in the Balkans—BalkanCryptSec 2014*, vol. LNCS 9024, pp. 15–30, Springer-Verlag 2015.
- [23] E. PASALIC. Probabilistic versus deterministic algebraic cryptanalysis - a performance comparison. *IEEE Trans. on Inform. Theory*, vol. 55, no. 11, pp. 2182–2191, 2009.

- [24] E. PASALIC. On guess and determine cryptanalysis of LFSR-based stream ciphers. *IEEE Trans. on Inform. Theory*, vol. 55, no. 7, pp. 3398–3406, 2009.
- [25] T. SIEGENTHALER. Decrypting a class of stream cipher using ciphertext only. *IEEE Trans. Comp.*, vol. C-34, no.1, pp.81–85, January 1985.
- [26] Y. WEI, E. PASALIC AND Y. HU. Guess and determinate attacks on filter generators–Revisited. *IEEE Trans. on Inform. Theory*, vol. 58, no. 4, pp. 2530–2539, 2012.
- [27] A. BRAEKEN, J. LANO, N. MENTENS, B. PRENEEL, AND I. VERBAUWHEDE. SFINKS: A synchronous stream cipher for restricted hardware environments. *eSTREAM, ECRYPT Stream Cipher Project*, Report 2005/026 (2005)

## Appendix

Since finding the preimage spaces  $S_{w^{t_i}}$  of the observed outputs  $w^{t_i}$  is the most important part, we give for clarity the description of a few initial steps:

**Step 1:** Let  $w^{t_1}$  denotes the first observed output so that the corresponding LFSR state at the tap positions is exactly the set  $\mathcal{I}_0 = \{l_1, l_2, \dots, l_n\}$ , so that  $s^{t_1} = (s_{l_1}^{t_1}, \dots, s_{l_n}^{t_1}) \stackrel{(4)}{=} \{l_1, \dots, l_n\}$ , i.e.,  $s^{t_1} = \mathcal{I}_0$ . Notice that the first observed output  $w^{t_1}$  does not necessarily need to correspond to the set  $\mathcal{I}_0$ , though (for simplicity) we assume this is the case. A preimage space which corresponds to the first observed output  $w^{t_1}$  always has the size  $2^{n-m}$ , i.e.,  $|S_{w^{t_1}}| = 2^{n-m}$ .

**Step 2:** Taking the second output  $w^{t_2}$  at distance  $\sigma_1$  from  $w^{t_1}$  (thus  $t_2 = t_1 + \sigma_1$ ), we are able to identify and calculate the number of repeated bits (equations) at the time instance  $t_2$ . Using the notation above, the set  $\mathcal{I}_1^*$  of these bits is given by the intersection:

$$\mathcal{I}_1^* = s^{t_1} \cap s^{t_2} = s^{t_1} \cap \{s^{t_1 + \sigma_1}\} = \mathcal{I}_0 \cap \{l_1 + \sigma_1, \dots, l_n + \sigma_1\},$$

where  $s^{t_2} = \{s^{t_1 + \sigma_1}\} \stackrel{\text{def}}{=} (s_{l_1 + \sigma_1}^{t_1 + \sigma_1}, s_{l_2 + \sigma_1}^{t_1 + \sigma_1}, \dots, s_{l_n + \sigma_1}^{t_1 + \sigma_1}) \stackrel{(4)}{=} \{l_1 + \sigma_1, \dots, l_n + \sigma_1\}$ , i.e.,  $s^{t_2} = \{l_1 + \sigma_1, \dots, l_n + \sigma_1\}$  is the LFSR state at tap positions at time instance  $t_2$ . Denoting the cardinality of  $\mathcal{I}_1^* = s^{t_1} \cap s^{t_2}$  by  $q_1$ , i.e.,  $q_1 = \#\mathcal{I}_1^*$ , the cardinality of the preimage space corresponding to the output  $w^{t_2}$  is given as  $|S_{w^{t_2}}| = 2^{n-m-q_1}$ .

This process is then continued using the sampling distances  $\sigma_2, \dots, \sigma_{c^*}$  until the condition  $nc^* - R^* > L$  is satisfied, where the total number of repeated equations over  $c^*$  observed outputs is  $R^* = \sum_{k=1}^{c^*-1} q_k$ . Note that the number of repeated equations corresponding to the first output is 0, since the corresponding LFSR state  $s^{t_1}$  is the starting one. Therefore the sum goes to  $c^* - 1$ .